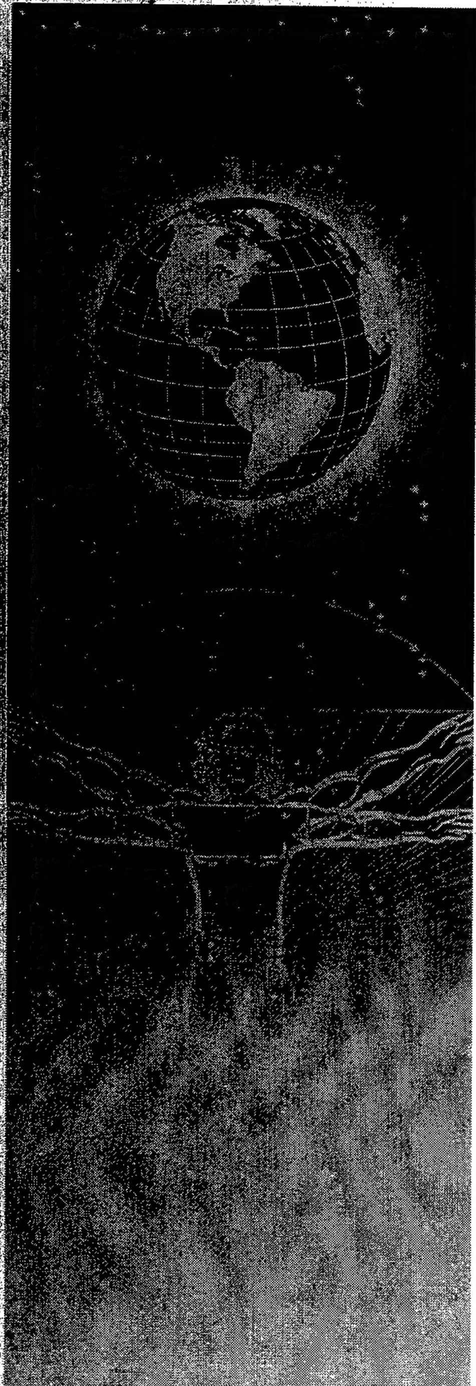
AFRL-HE-WP-TR-2003-0149



UNITED STATES AIR FORCE RESEARCH LABORATORY

AN EXAMINATION OF COMPLEX HUMAN-MACHINE SYSTEM PERFORMANCE UNDER MULTIPLE LEVELS AND STAGES OF AUTOMATION

Scott M. Galster

HUMAN EFFECTIVENESS DIRECTORATE
CREW SYSTEM INTERFACE DIVISION
WRIGHT-PATTERSON AFB OH 45433-7022

OCTOBER 2003

INTERIM REPORT FOR THE PERIOD JULY 2000 TO SEPTEMBER 2003

Human Effectiveness Directorate
Crew System Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

NOTICES

DISCLAIMER

This Technical Report is published as received and has not been edited by the Air Force Research Laboratory, Human Effectiveness Directorate.

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-2003-0149

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

//Signed//

MARIS M. VIKMANIS
Chief, Warfighter Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE						Form Approved OMB No. 074-0188
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**						
1. REPORT DATE (DD-MMM-YYYY) October 2003		**2. REPORT TYPE** Interim Report				**3. DATES COVERED (From – To)** July 2000 – September 2003
4. TITLE AND SUBTITLE An Examination of Complex Human-Machine System Performance Under Multiple Levels and Stages of Automation						**5a. CONTRACT NUMBER**
						5b. GRANT NUMBER
						5c. PROGRAM ELEMENT NUMBER 62202F
6. AUTHOR(S) Scott M. Galster						**5d. PROJECT NUMBER** 7184
						5e. TASK NUMBER 08
						5f. WORKUNIT NUMBER 65
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)						**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory Human Effectiveness Directorate Crew System Interface Division Air Force Materiel Command Wright-Patterson AFB OH 45433-7022						**10. SPONSOR / MONITOR'S ACRONYM** AFRL-HE-WP-TR-2003-0149
						11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.						
13. SUPPLEMENTARY NOTES						
14. ABSTRACT The introduction of automation into highly complex systems has occurred under several guiding principles. The application of these principles has often resulted in tenuous interactions with regard to human performance within complex systems. With advances in technology increasing, it is no longer applicable to look at single automated tools but rather at how several automated tools fit together and affect system performance. A common framework utilizing a model of human interaction with automation based on simple human information-processing stages was used in the design and analysis of 4 experiments. The first 3 experiments utilized a visual search paradigm and varied the stage the automation was present and the reliability of the automation that was used. For these studies, the automation that helped the operator locate the potential target demonstrated a clear advantage over automation that recommended a course of action when the automation was perfectly reliable. The 4th study examined all of the possible combinations of manual & automated aiding for the 4 stages in an air-to-ground search and destroy mission that was carried out in a high fidelity combat flight simulator. By utilizing separate stage metrics, it was demonstrated that the automation in 1 stage influenced performance in subsequent stages and throughout the entire mission.						
15. SUBJECT TERMS Automation, Human-Interaction, Automation Model, Complex Systems, Human Performance						
16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON:** Scott M. Galster, PhD	
a. REPORT UNCLAS	**b. ABSTRACT** UNCLAS	**c. THIS PAGE** UNCLAS	UNLIMITED	204	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-8737	

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39-18

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The introduction of automation into highly complex systems has occurred under several guiding principles. The application of these principles has often resulted in tenuous interactions with regard to human performance within complex systems. With advances in technology increasing at an exponential rate it is no longer applicable to look at single automated tools but rather at how several automated tools fit together and affect system performance. Automation can change the nature of the demands on the operator and produce subsequent changes in performance not seen when automation is absent. Problems in human-automation interaction have included unbalanced workload, reduced system awareness, decision bias, mistrust, over-reliance, complacency, and reduced manual skills. Further, these problems can be exacerbated when the automation is less than perfectly reliable.

A common framework utilizing a model of human interaction with automation based on simple human information-processing stages was used in the design and analysis of four experiments. The model was used for tasks that varied in complexity and the amount of automation that was available to the operator. The first three experiments utilized a visual search paradigm and varied the stage the automation was present and the reliability of the automation that was used. For these studies, the automation that helped the operator locate the potential target (information automation) demonstrated a clear advantage over automation that recommended a course of action (decision-aiding automation) when the automation was perfectly reliable. Costs associated with unreliable automation generally were greater for the information automation stage, which was not congruent with the results of previous research.

The fourth study examined all of the possible combinations of manual and automated aiding for the four stages in an air-to-ground search and destroy mission that was carried out in a high fidelity combat flight simulator. By utilizing separate stage metrics, it was demonstrated that the automation in one stage influenced performance in subsequent stages and throughout the entire mission. These benefits were apparent in the primary task performance and the subjective ratings of mental workload, situation awareness, and trust in the automation.

ACKNOWLEDGEMENTS

I consider it an honor and privilege to dedicate this work to my family and friends who have shown non-wavering support throughout this endeavor. I thank my parents for their steadfast belief that it was always possible and I would like to single out the support I have received from my wife Mary as well as Jon and Jen, without whom this would not have been possible.

I would also like to express a deep sense of gratitude to my advisor, mentor, and friend Raja, who epitomizes the role of being a teacher. Further, I thank Todd for being approachable and providing critical but consistently accurate feedback on experimental methodology and previous drafts. Also, I thank Marc for his insights, and efforts to make this journey pleasant.

I owe a great debt of gratitude to Matt Middendorf for his effort and ingenuity: people like him seem to make the world go around. Becky Brown made it all happen, and various people at my work made contributions to allow it to happen (Robert Bolia, Grant McMillan, Mike Haas, Mike Vidulich, Mark Draper, Gloria Calhoun, Glen Geisen, Mike Poole, Jeff Collier, Merry Roe, Heath Ruff and many more), I thank you all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3-D SART	Three Dimensional Situation Awareness Rating Technique
A/G	Air-to-ground
AD	Attack Display
ANOVA	Analysis of Variance
APC	Armored Personnel Carrier
ASRS	Aviation Safety Reporting System
BCIS	Battlefield Combat Identification System
CMS	Countermeasures Management Switch
DA	Decision-Aiding
DCom	Distributed Communication
DIS	Distributed Interactive Simulation
DD	Defensive Display
DMS	Display Management Switch
FEBA	Forward Edge of the Battle Area
FOR	Field-of-regard
FOV	Field-of-view
HDD	Head-Down Display
HOTAS	Hands On Throttle and Stick
HUD	Head-Up Display
IA	Information Automation
IEEE	Institute of Electrical and Electronic Engineers, Inc.
M	Manual
MAT-B	Multi-Attribute Task Battery
ModSAF	Modular Semi-Automated Forces
NASA-TLX	National Aeronautics and Space Administration Task Load Index
NTSB	National Transportation Safety Board
PDU	Protocol Data Unit
RBGM	Real Beam Ground Map
RWR	Radar Warning Receiver
SA (1)	Situation Awareness
SA (2)	Surface to Air defense system
SAM	Surface-to-Air Missile
SAR	Synthetic Aperture Radar
SD	Situation Display
SIRE	Synthesized Immersion Research Environment
THAAD	Theater High Altitude Area Defense
TMS	Target Management Switch

INTRODUCTION

Part 1 - Automation Models

The introduction of computers and subsequent advances in technologies have dramatically changed the nature of many work environments. System designers, aware of the potential increases in productivity, began to utilize computers to free human operators from repetitive, and often ill-suited tasks. As the computational power of computers increased, designers found more tasks amenable to computer oversight. This created a new dimension in the relationship between humans and machines. Most often, computers were used to automate tasks previously carried out by humans. Automation is defined by Parasuraman and Riley (1997) as a device or system (usually a computer) that accomplishes (partially or fully) a function that was previously carried out (partially or fully) by a human operator. The decision to automate a task is often compelled by the promise of potential benefits afforded by the automation. Automated systems have provided a myriad of benefits including increased productivity and efficiency and a reduction in many of the costs (financial and error induced) associated with operating complex systems. Automated tools have allowed the operation of more complex systems than would otherwise be possible without automation (Woods, 1996). These benefits have been obtained in the aviation domain (Billings, 1991; Wiener, 1988), in maritime operations (Lee & Sanquist, 1996), and in air traffic control (Wickens, Mavor, & McGee, 1997). Furthermore, automation has been widely used in areas such as aviation, ground and sea transportation systems, process control and manufacturing plants as well as in the medical domain (Parasuraman & Mouloua, 1996; Sheridan, 2002). Increases in speed and sophistication coupled with decreases in size and cost will assuredly promote the introduction of automation to many other domains in work and everyday life.

The integration of automated tools into highly complex systems has created a need to examine the nature of the interaction between humans and the automated tools that they use. The examination of human performance in systems that utilize automation has become widespread. Investigations of human interaction with automation have revealed that

automation does not always function in the way intended by designers and, moreover, can produce deleterious performance effects (Bainbridge, 1983; Billings, 1997; Billings & Woods, 1994; Parasuraman & Riley, 1997; Sarter & Woods, 1995; Wiener & Curry, 1980; Woods, 1996). Automation can change the nature of the demands on the operator and produce subsequent changes in performance not seen when automation is absent. A review of human performance costs of automation list possible changes in the mental workload for the operator, an increase in the monitoring demands, and a decrease of the monitoring efficiency of the operator as costs contributing to poor performance (Parasuraman & Riley, 1997; Sheridan, 2002). A reduction in skill due to lack of use and a reduction in the situation awareness of the operator have also been identified as potential costs of automation usage.

Advanced automated tools are used most frequently in the control of complex systems including aviation, nuclear power control rooms and ocean-going vessels. It would be inconceivable for an operator to perform competently within these and similar systems without the use of automated tools to help maintain the balance of the system. The goal of these systems is to operate in a safe, efficient, and profitable manner. In order to do so, the design engineers need to decide how much of a particular operation or task to automate. On one hand, the human cannot possibly do all of the work; on the other hand very few systems will operate flawlessly in a fully automated state.

The quest to find the balance between human and automation control is not new and is associated with the long held (and often criticized) notion of function allocation. Fitts (1951) addressed the function allocation question before there was widespread usage of computers in his well known comparison of the relative capabilities of humans and machines. At that time computers served primarily as processors of raw data. Fitts (1951) created a list of allocation strategies dependent on the task that was being performed. If the machine was better at performing a particular task then the machine was allocated that task. Examples of machine superiority (at that time) included the ability to respond quickly to control signals, perform repetitive and routine tasks, and the ability to reason deductively and process information in a parallel manner. The areas where humans were thought to be superior included the perception of patterns in light or sound, the ability to improvise and reason inductively, and the ability to exercise judgment. Consequently, these tasks should have been allocated to humans. Notably, this list was developed for the allocation of the task in the

design phase of the complex system. As such, once the allocation was made it was not changed, an example of a static allocation of function.

It should be mentioned that the allocation of responsibilities was addressed relatively early. Jordan (1963), an early critic of the Fitts' list approach to function allocation, suggested that a direct comparison between people and machines was dubious when considering what to delegate. Rather, Jordan suggested that people and machines should be considered complementary and allocation decisions should be made with this in mind.

Levels of Automation

A major change addressing the allocation of functions was the realization that automation should not be considered as an either/or (binary) entity. Instead, automation could be granted more or less authority depending on the nature of the task or situation. Moreover, the automation level could be capable of being changed adaptively during system operations as the situation dictated; this approach is typically termed adaptive automation (Parasuraman, Bahri, Deaton, Morrison, & Barnes, 1992) or dynamic function allocation (Hancock and Chignell, 1987). Several authors have proposed different schemes or models to quantify the levels of automation (Billings, 1991; Rouse & Rouse, 1983; Sheridan, 1980). Table 1 outlines the scale of levels of automation outlined by Sheridan and Verplank (1978) for decision and control actions.

Automated tools used for decision and control actions can function at specific levels according to this methodology. For example, an automated collision avoidance system may operate at level three normally and advise pilots of several options that may be available to avoid conflicting aircraft. This tool could also operate at level seven if programmed to do so and would execute an evasive maneuver first then let the pilot know what it had commanded. This automated collision avoidance tool could be set at a particular level depending on its operating characteristics (e.g., time to collision, traffic density, airline management standards, etc.) indicative of a static allocation. Conversely, an adaptive allocation may exist which takes into account pilot and aircraft performance envelopes. Under either allocation strategy, the level of automation can be quantified according to task responsibility.

HIGH	10.	The computer decides everything and acts autonomously, ignoring the human
	9.	Informs the human only if it, the computer, decides to
	8.	Informs the human only if asked
	7.	Executes automatically, then necessarily informs the human
	6.	Allows the human a restricted time to veto before automatic execution
	5.	Executes the suggestion if the human approves
	4.	Suggests one alternative
	3.	Narrows the selection down to a few alternatives
	2.	The computer offers a complete set of decision/action alternatives
LOW	1.	The computer offers no assistance: the human must make all decisions and actions

Table 1. Levels of automation for decision and control actions.

Current Trends in Automation Modeling

As pointed out previously, the levels of automation depicted in Table 1 were devised for decision and control actions. This corresponds to the <u>output</u> functions of a system. There are, however, automated tools which operate at the input side of system functionality, those that acquire and analyze information but are not intended to be associated with decision making or action selection. To rectify the uni-dimensional approach of previous models, Parasuraman, Sheridan, and Wickens (2000) proposed a 2-dimensional model for the types and levels of human-interaction with automation. The model is based on a four-stage simplified human information-processing or perception-action model (see Figure 1). While the authors concede that this approach is elementary in its approach to human cognition it is important to realize that the architecture it provides can be instrumental in understanding the nature of the relationships between stages and levels of automation.

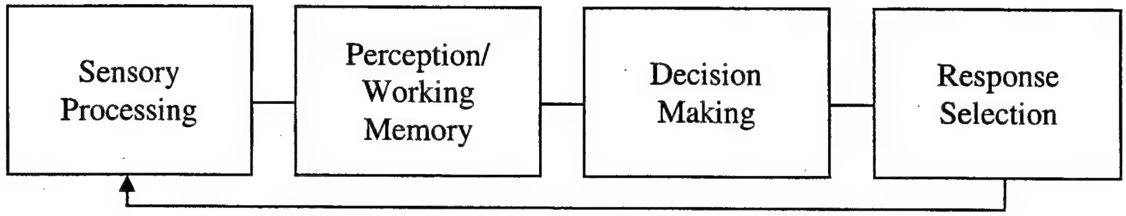


Figure 1. Simple four-stage model of human information processing.

In human information processing, examples of first stage elements include raw sensory data from the external environment, and items that drive selective attention mechanisms. Corresponding automation stage one elements may include raw radar data or partial- pattern recognition items in intelligent agent software. Full conscious perception and manipulation of data in working memory (e.g., rehearsal, integration, etc.) occur in the second stage of human information processing (Baddeley, 1996). Examples of automation elements at this stage include the use of integrated displays, systems that present information in multiple modalities, and tools that facilitate the analysis and presentation of data. It should be noted that the first two stages, referring to automation of human information acquisition and analysis, occur prior to the point of decision (Parasuraman et al., 2000). The third stage represents the decision that is made after due consideration has been given to the veracity of the information from the previous stages. This denotes the beginning of the output side of the process. The fourth stage in this process encompasses the effectuation of an action congruent with the decision option.

Parasuraman et al. (2000) stress that they are not attempting to describe, debate, or propose a specific human information-processing model. Rather, they use the information-processing model as an outline to describe a model of automated system functions. In their model (see Figure 2), automation can be assigned to a stage, pursuant to the function that it performs in the system. Accordingly, they name their stages information acquisition (acquisition), information analysis (analysis), decision and action selection (decision), and action implementation (action), respectively. Furthermore, they note that information automation may include both the acquisition and analysis stages jointly.

Complex systems can include automation across all stages at differing levels. System B in Figure 2, for example, has a relatively high level of automation across all four

dimensions of automation. System A, however, has a high information acquisition level followed by a relatively low level of automation in the remaining stages. The level in one stage do not necessarily conform to levels of automation in other stages. The list provided by Sheridan and Verplank (1978) was designed for the decision and action selection stage; thus it is not prudent to apply this list to all of the different stages. Each stage has a unique continuum that will allow an expression of the varying degrees of automation.

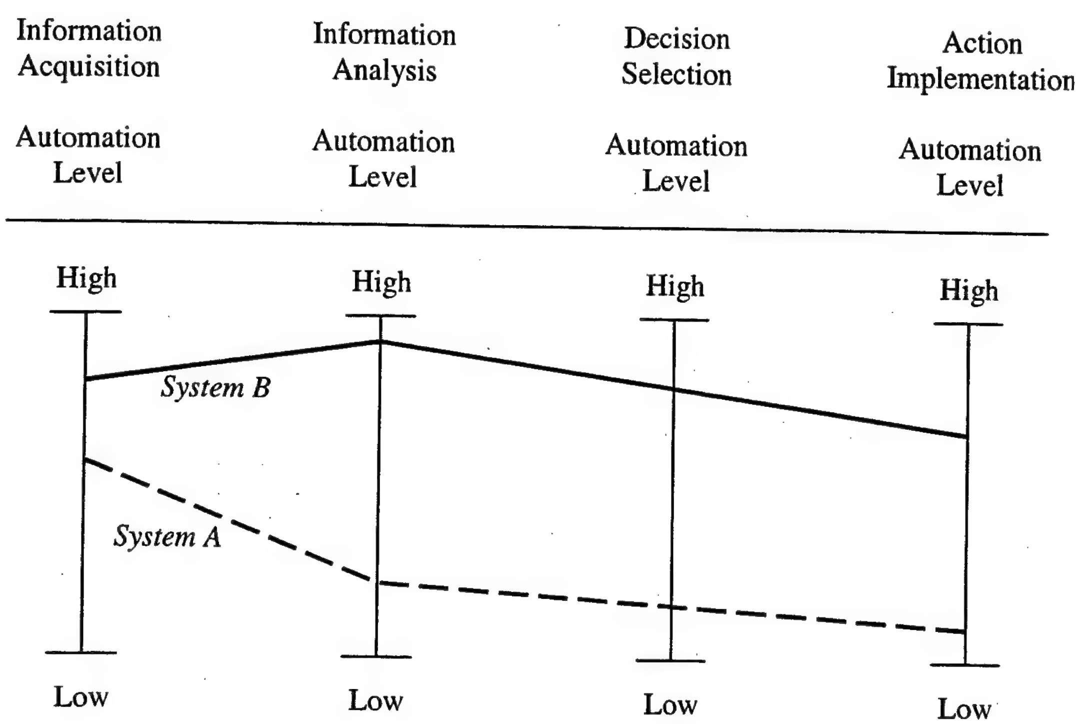


Figure 2. Levels of automation for independent functions of information acquisition, information analysis, decision selection, and action implementation. Examples of systems with different levels of automation across functional dimensions are also shown.

Evaluation Criteria

The model proposed by Parasuraman et al. (2000) is intended to be used as a criterion-based design tool. As such, iterative testing is needed to determine the automation levels appropriate for each stage of system functionality. To accomplish this, the authors

proposed a set of evaluative criteria that serve to guide the design community. By applying the outlined principles, a range of automation levels can be reached that defines the upper and lower bounds of automation for each stage. The testing methodology proposed is hierarchical and recursive. Initially, primary evaluative criteria are used to assess human performance given a set of predetermined automation parameters. Criteria at this stage determine whether a combined set of automation levels offer an enhanced human/system performance or, conversely, provide a decrement or no change in performance. After completing the performance evaluation, a second set of evaluative criteria is used to test consequences related to variables indigenous to the use of automation at each stage.

Primary Evaluative Criteria

Prior research in human performance has shown that both benefits and costs can be associated with the introduction of automation (Parasuraman, 1993; Parasuraman & Mouloua, 1996; Parasuraman, Mouloua, & Hilburn, 1999; Parasuraman & Riley, 1997; Wiener, 1988; Woods, 1996). Of particular concern are the often unanticipated costs that accompany the use of automation from a human/system performance perspective. Consequences of these costs have ranged from temporary confusion (Sarter & Woods, 1995) to the loss of human life (NTSB, 1973, 1986; Stein, 1983). Research into the causes of these costs has identified increases in mental workload, a loss of situation awareness, skill degradation, and automation-induced complacency as potential contributing factors (Kessel & Wickens, 1982; Knapp & Vardman, 1991; Parasuraman, Molloy, & Singh, 1996; Parasuraman et al., 2000; Sarter, Woods & Billings, 1997). Operator trust and acceptance have also been identified as potential problematic factors (Lee & Moray, 1992, 1994; Masalonis, 2000; Muir, 1988; Sheridan, 1988; Wickens, 1994). While this list is not exhaustive, it points to the types of performance measurements and matrices that should be used when performing a primary evaluation of automated systems. Implicit in this stage of evaluation are objective measurements of performance that reveal system goal attainment. For example, an evaluation of an automated target acquisition tool should examine an operator's ability to identify a target faster and/or with greater accuracy when automation is present compared to when this task is performed manually. Conducting a primary evaluation

is essential in determining the level of automation that should initially be applied at each stage.

Secondary Evaluative Criteria

Once it has been shown that system performance can be increased without corresponding decreases in human performance, it is prudent to examine the effects of problems associated with the automation itself. Parasuraman et al. (2000) point out that the researcher/designer must evaluate automation reliability because it often determines the level of trust and hence, use of the automation. Furthermore, trust can be calibrated more or less rapidly depending on the stage at which the automation is unreliable and the amount of feedback regarding the performance of the automation (Merlo, Wickens, & Yeh, 2000; Wickens, Conejo, & Gempler, 1999). Another method of evaluation centers on the costs associated with decision/action outcomes. There is always a certain amount of risk involved in the use of automation, primarily at higher levels, for decision selection and action implementation. As defined by Parasuraman et al. (2000), the amount of risk is the cost of an error multiplied by the probability of that error. Determining the acceptable risk is the key to choosing the appropriate level of automation. Low-risk applications are amenable to higher levels of automation while high-risk applications, in general, should conform to lower levels of automation.

Evaluating the Model

The Parasuraman et al. (2000) model of human-interaction with automation does not prescribe the levels of automation that should be implemented in the system design process. Rather, it serves as an evaluation tool that offers a more complete and objective approach to automation system design compared to approaches that are based on technical capability or other considerations alone. Obviously, there are many different types of automated systems, each with their own performance objectives and level of specificity. Furthermore, each potential automated system will have a greater or lesser degree or emphasis on each stage of the model. In order to gauge the utility and applicability of the model, the approach must be used in a wide variety of automated systems. One measure of applicability is scalability; the

model must be able to address automated systems that differ in scope. If the model is robust, it should be able to scale between those systems that may be looking at tasks that can be achieved in seconds and those that will take much longer. For example, the four stages of information processing can be equally applied to a basic visual search task and an entire war effort. In order for the model to be considered robust it must be able to address the human-interaction with automation at extremes of temporal scale and complexity, e.g., from a simple satellite photographic identification task to be carried out at leisure, to a battlefield command and control situation requiring decisions to be carried out under severe time stress. In addition, task domains should not bind the model. The model should function equally well in the evaluation of an automated system in an aviation domain as it does in a medical or manufacturing domain.

Although this effort has several research goals, one of the goals will be to evaluate the model's ability to scale effectively and cross domains. The evaluation will be accomplished in two ways. First, the model will be applied to a basic time-limited visual search task that incorporates automation at various stages to help participants identify simulated targets and execute appropriate responses. Second, the model will be applied to an air-to-ground search and destroy mission in a simulated military aviation domain. These applications will test the model's ability to scale effectively and cross domains. The scale in the visual search task is small, 2.5s or less, while the scale of the search and destroy task is several minutes. The domains are also different. The visual search task is basic and requires participants who are taken from a general subject pool with no previous experience required. The search and destroy mission was complex and dynamic. The requirements for participation were much more stringent; only pilots with combat military experience could participate.

The present studies were developed, in part, to evaluate the model and its ability to scale temporally and with regard to task complexity. It is hypothesized that the model will be able to scale appropriately due to its reliance on the basic information-processing model. All tasks, simple and/or complex, require some amount of information processing. Thus the information processing nature of the Parasuraman et al. (2000) model should give it the flexibility to scale accordingly. It should be noted that there is not a prediction being made about the level of automation that is most appropriate to use, rather, the prediction is that if an automated aid is available the model can be used to evaluate it effectively.

The metrics used to determine model effectiveness are often vague. In this effort, if the model reliably provides information regarding the use of automation in one or more of the stages it is functioning in an effective and appropriate manner. The model was originally developed to be used as a tool that would facilitate the introduction of automated aids in system planning and design. Ideally, the effectiveness of the model would be evaluated against the success of a system that used it to determine the appropriate level(s) of automation in the design process. It can also be used, as it is here, as a research tool that allows the examination of system performance when automation is implemented at different stages and to different degrees. As a research tool, this or any other model can be evaluated by the elicitation of useful information that it provides over and above what could be obtained by not using the model.

Part 2 - Empirical Studies

While there is a solid base of research and technical information on automation, there is a small but emerging base of research on human capabilities and interactions with automated systems. Within that base of research there are still fewer examples of cases where human performance was examined as a function of the stage in which the automation was present. Below is a review of the literature that has examined human performance in automated systems as a function of the stage, or stages, during which the automation was present.

One of the primary reasons for including automation, assuming a human-centered approach, is to increase overall system efficiency and performance. There are examples that automation does in fact improve performance. Wickens and Dixon (2002) demonstrated that autopilots fly airplanes with greater accuracy than when the aircraft is controlled manually. Yeh and Wickens (2001) produced results showing that automated cuing improved visual searches over searches where there was not an automated cue. In contrast, benefits of automation do not always occur. Rovira, McGarry, and Parasuraman (2002) suggested that benefits are tied to task complexity. If the manual task is relatively easy, automating it does not produce an increase in performance. Further, there is some evidence that performance improvements may be made only in difficult, high workload task environments (Merlo et al., 2000) or under temporally demanding task situations (Muthard & Wickens, 2001).

Reliability of the Automation

A majority of the studies conducted also manipulated the reliability level of the automation as a separate factor. It is imperative to note the importance of the reliability of the automation in the examination of system performance. Reliability levels can affect both the primary evaluative criteria and the secondary evaluative criteria when the Parasuraman et al. (2000) model is applied. An example of one of the primary evaluative concerns (complacency) is given below. This is not to say that other primary evaluative concerns (i.e.,

mental workload, situation awareness, skill degradation) are not affected by reliability level of the automation. Rather, it illustrates that that the primary evaluative criteria are based on empirical observations over many studies.

Automation Induced Complacency

Automation reliability levels have been shown to induce over-reliance on the automation or automation-induced complacency. Complacency is not a new concept and has even been linked as a contributing factor in many aviation accidents (Hurst & Hurst, 1982). As systems have become more automated, complacency has been identified as a causal factor to a greater degree. Evidence of this is the inclusion of complacency as a behavioral coding category used by the Aviation Safety Reporting System (ASRS) to classify aviation incidents and accidents (Sumwalt, Morrison, Watson, & Taube, 1997). There are differing opinions as to the definition of complacency. Billings, Lauber, Funkhouser, Lyman and Huff (1976), writing on the ASRS, defined complacency as "self-satisfaction which may result in non-vigilance based on an unjustified assumption of satisfactory system state" (p. 23). Another definition proposed by Wiener (1981) defines complacency as "a psychological state characterized by a low index of suspicion" (p. 117). It is generally agreed that, however defined, complacency is a potential pathogen in aviation incidents and accidents.

Thackray and Touchstone (1989) performed what is believed to be the first empirical test of complacency using a very simple air traffic control task. Participants were asked to detect aircraft conflicts during a two-hour monitoring task with or without an automated aid that provided a message indicating a conflict situation. Their hypothesis was that the participants would detect fewer conflicts with the automated aid than they would while not using the aid. The automated aid failed to detect a conflict once early in the test session and once late in the test session. The results indicated that the participants were equally efficient at conflict detection regardless of the automation condition. Thackray and Touchstone surmised that two possible reasons existed for their failure to find complacent behavior. First, the test session (2 hrs.) might have been too short for complacent behavior to develop and secondly the participants were engaged in only a single task, monitoring for a conflict.

Parasuraman et al. (1993) suggested that a lack of confirmation for a complacency effect by Thackray and Touchstone (1989) may be attributable, in part, to the task their participants were performing. Information provided in ASRS reports (Billings et al., 1976; Mosier, Skitka, & Korte, 1994) indicated that many of the monitoring failures experienced by crewmembers occurred under multiple task conditions, very common in the cockpit environment. Thackray and Touchstone however used only a single monitoring task. Parasuraman et al. reasoned that a complacency effect might be shown if the operator was engaged in a multiple task environment. They tested non-pilot participants using a modified version (Parasuraman, Bahri, & Molloy, 1991) of the Multi-Attribute Task Battery (MAT-B) developed by Comstock and Arnegard (1992). The MAT-B consists of three tasks: a two-dimensional compensatory tracking task, an engine-monitoring task and a fuel resource task.

In the multi-task environment, participants were asked to perform the tracking and fuel resource tasks manually while an automation routine managed the engine-monitoring task. The reliability level of the automation was manipulated (87.5% vs. 56.25%) by changing the failure rate of the automation routine to detect an engine deviation. Another factor in this study was the consistency of the automation reliability. The automation reliability levels were either constant at the aforementioned rates for two groups or they alternated every 10 minutes, counterbalanced, for the remaining two groups. The results indicated that 72% of engine malfunctions were detected in the manual condition while only 37% and 28% of failures were detected in the constant low-reliability and constant high-reliability conditions, respectively. The variable reliability groups however, performed well, detecting an average of 82% of the automation failures. Thus, Parasuraman et al. (1993) found a complacency effect for the constant reliability groups under multi-task conditions. To support their claim that complacency would emerge under multi-task environments, Parasuraman et al. conducted a second study in which the monitoring automation routine was the only task performed. The results indicated that the detection of failures under manual and automation conditions was equally good (98%) confirming the lack of evidence found by Thackray and Touchstone (1989) for complacency in a single task environment.

Subsequent investigations have found evidence for automation-induced complacency when pilots were used as participants (Parasuraman, Mouloua, & Molloy, 1994) and when only one automation failure was present (Molloy & Parasuraman, 1996). Complacency

effects were also found when the display was moved to a central location (Singh, Molloy, & Parasuraman, 1997) and when the monitoring task was superimposed on the tracking task (Duley, Westerman, Molloy, & Parasuraman, 1997). Furthermore, Farrell and Lewandowsky (2000) have presented a connectionist model of complacency that suggests that automation-induced complacency is due in part to divergent operator learning processes for monitoring under automation and manual control.

Empirical Automation Stage Based Studies

There are two types of empirically based studies, those that examined automation at one stage, such as cueing studies, and those that examined automation at more than one stage. In a study on automated cueing, Wickens et al. (1999) found that pilot detection performance decreased when a cue incorrectly guided attention away from the target even when the pilots knew the cue was not totally reliable. In another cueing study, Yeh, Wickens, and Seagull (1999) found that operators did not effectively pay attention to un-cued areas of a display in a ground target detection task. This overtrust in automation has also been replicated in tasks where the automation directs the pilot's attention to system failures (Mosier, Skitka, Heers, & Burdick, 1998); and in rotorcraft hazard cueing (Davison & Wickens, 2001).

In an attempt to look at differential performance effects by stage of automation, Crocoll and Coury (1990) examined decision-aiding performance when operators were given status, recommendation, or status and recommendation cues in an aircraft identification task. The first two of these conditions can be associated with the information analysis and decision selection stages of automation in the subsequently developed Parasuraman et al. (2000) model. Operators were required to visually identify aircraft as being hostile, friendly or unknown and then choose a fire or no fire response in accordance with stated rules of engagement. The "tight" rule of engagement allowed the operator to fire only upon hostile aircraft while the "free" rule of engagement allowed firing upon hostile and unknown aircraft. During the first three sessions, participants learned how to identify 10 friendly and 10 hostile aircraft, identify unknown aircraft types, and apply the rules of engagement

criteria. In the fourth session, the data collection session, participants were divided into four groups and tested on their ability to choose the correct engagement decision. The first group was the control group and received no aiding. The second, third, and fourth groups received status only, recommendation only, or status and recommendation aiding, respectively. The decision aiding was reliable 96% of the time when the automation was present. The percent of correct engagement decisions made and the response times were recorded. It was unclear if the trials were time limited or if they continued until the participant responded.

The percent of correct engagement decisions was greater than 96% for all conditions and did not show a significant difference between the automated and control conditions. The response times significantly improved when the automation was present compared to the non-aided control group but there was not a significant difference between the three aided conditions. Crocoll and Coury (1990) decided to examine the performance on the automation-aided trials to see if there was a difference when the aid was unreliable (8 of the 200 trials for each group). They found that the group that received the status only aid responded correctly 95% of the time while the status and recommendation, and the recommendation only groups responded correctly 86% and 80% of the time respectively. The data indicated that there was a greater cost when the recommendation aiding was present compared to the status only or the status and recommendation aiding conditions. Crocoll and Coury surmised that participants who were provided a recommendation decision aid blindly followed that aid compared to the participants who received the status only or status and recommendation decision aiding.

Sarter and Schroeder (2001) conducted a study comparing pilot performance during escalating in-flight icing conditions using two types of decision-aids during simulated flight. The first decision aid in their study presented icing information (status display) and the other decision-aid recommended actions to mediate the icing condition (command display). They demonstrated that imperfect automation led to reduced performance while using the decision aiding (command display) over both the status display and the baseline condition where no automation was present. This result is consistent with the suggestion that the negative effects of unreliable automation in the decision stage may be more pronounced than the information analysis stage (Parasuraman et al. 2000).

Rovira, McGarry, and Parasuraman (2002) also found a greater cost in performance when the decision-aiding automation was unreliable compared to when the information analysis stage was unreliable in a sensor-to-shooter task. These effects generalized across three different forms of decision automation. Furthermore, they found that this performance decrement dropped below manual performance as measured by the percentage of correct detections in a command and control task. In addition, they included varying reliability rates (80% vs. 60%) and noted that there was a greater cost in the decision-aiding stage than in the information analysis stage. This cost was greater in the higher reliability condition compared to the lower reliability condition, consistent with the findings on automation complacency reviewed earlier (Parasuraman et al., 1993). McGarry, Rovira, and Parasuraman (in press) found similar results but also noted that the findings applied to tasks that were longer in duration than the original sensor-to-shooter task that was reported by Rovira, McGarry et al. (2002).

A similar pattern of results was obtained in a multi-task environment using the MAT battery (Rovira, Zinni, & Parasuraman, 2002). There was a general decline in performance when the automation was unreliable over when it was reliable. Also, there was a differential performance decrement for the unreliable automation conditions depending on what stage the automation was employed. There was a greater drop in performance when the automation was employed in the decision-aiding stage over the information analysis stage. Further, the results indicated that the higher reliability rate induced a greater cost in detections, again indicating a complacency effect that was similar to that found by Parasuraman et al. (1993).

Secondary task performance has also been examined as a factor when automation is included or excluded in task designs. Metzger and Parasuraman (2001) found that air traffic controllers performed the secondary task of updating flight progress more accurately but slower under automated conditions as compared to manual conditions. Significantly improved detection rates and response times were also found in the automated conditions for the primary task of detecting potential conflicts between aircraft in a Free Flight environment. Lorenz, Di Nocera, and Parasuraman (2002) found a similar secondary task benefit under automated conditions as compared to manual conditions. Participants responded faster to alarms during a simulated spaceflight operation.

Rationale

There were several issues that emerged in the review of stage specific automation usage. Foremost, one can question whether introducing automation is actually beneficial as determined by system performance gains. Secondly, are the potential benefits attenuated by the stage in which the automation is introduced, the workload level imposed on the operator, the temporal compression for completing the required tasks, or the complexity of the task assignments? Finally, what effect does unreliable automation have on human/system performance and does this change depending on where the automation failure occurs in the information-processing cycle? These issues are worth examining because, as noted previously, while there is a small amount of literature on the effects of some of these factors, most of the previous work has been conducted in the context of either a binary (automation on/off) or unidimensional (level of decision automation) concept of automation. In contrast, the present work examined human interaction with automation in the context of a multidimensional (stage) concept of automation, as specified in the Parasuraman et al. (2000) model.

The present research effort is composed of four individual studies. The first three utilized a basic visual search task while the fourth was conducted in a complex and dynamic high-fidelity simulator. The basic visual search task was utilized across the first three studies to ensure a common testing environment. To date, a common testing environment has not been used to explore incremental changes in the use of automation by the stage it is implemented. Utilizing this common environment, the first study examined the differences in target detection and response times between manual and automated cueing conditions. The automated cuing condition represents the fusion of the information acquisition and analysis stages. As pointed out by Parasuraman et al. (2000), these stages are commonly combined because they occur prior to the decision-making point and represent information automation. The number of distractors in the search area was manipulated to represent varying levels of workload. In this and every study that used this task environment, a response was required within 2500ms for the presence or absence of a target among the distractor set. The purpose of the first study was to; (a) evaluate the visual search cueing

platform (Yeh & Wickens, 2001); (b) apply a simplified human interaction with automation model (Parasuraman et al., 2002); and (c) use a simple task (Rovira, McGarry et al., 2002) in the evaluation of the benefits of automation in high and low workload conditions (Merlo et al., 2000) under considerable temporal constraints (Muthard & Wickens, 2001). Further, the reliability of the automated cue was manipulated so that cue validity effects could be examined (Wickens et al., 1999; Yeh, et al., 1999).

The second study included a recommendation cue similar to the one used in the study by Crocoll and Coury (1990). A higher distractor set size was also added to increase the variability of the workload. In addition to the manual, information automation, and decision-aiding automation conditions the latter two were combined and presented either together (co-located) or separately resulting in five automation conditions.

As Wickens and Xu (2002) have noted, automation reliability levels seem to influence human-system performance differently, depending on the stage of automation. The third study varied the reliability level of the automation as a between-groups factor. All other experimental factors from the previous study were unchanged except the condition where the combined information automation and decision-aiding cues that were presented separately was dropped. This study allowed for the examination of human-system performance differences as the reliability level was manipulated between stages, similar to the Crocoll and Coury (1990), Sarter and Schroeder (2001), Rovira, McGarry et al. (2002), and Rovira, Zinni et al. (2002) studies. These studies did not treat the reliability level of the automation as a between-subjects factor. By including this in the third study the potential human-system performance changes by stage can be examined as a function of the reliability level experienced by the operators.

The fourth and last study included in this effort was designed to evaluate the scalability of the Parasuraman et al. (2000) model. The scope of the task was changed from a relatively simple, time-limited target search task to a complex and dynamic air-to-ground search and destroy task. Each stage of the model was mapped onto a stage within the overall air-to-ground search and destroy mission. Each of the four stages of the mission could be performed manually or with the aid of automation. Accordingly, each participant received every combination of manual and automation aiding in every stage. In addition, the workload level of the mission was manipulated and treated as a separate factor. To date,

there has not been an examination of the complete set of possible manual/automated conditions. Furthermore, with the exception of Clamann, Wright, and Kaber (2002), an examination of all four stages of the Parasuraman et al. model within one task environment has not been conducted. It should be noted that the Clamann et al. study treated the stage of automation as a between-subjects factor in the examination of the efficacy of utilizing adaptive automation between psychomotor tasks and cognitive tasks.

The search and destroy mission was the primary task for the participants. In addition, secondary tasks were included in the search and destroy mission as they have been shown to be sensitive to automation manipulations (Metzger & Parasuraman, 2001; Lorenz et al., 2002).

While these studies address these issues at a "basic science" level there are real world implications that make them worth examining. For example, the potential for high fratricide rates in combat has led to the development of automated aids such as the Battlefield Combat Identification System (BCIS), which was designed as a decision aid for the identification of friendly troops by armor gunners. This system sends a microwave signal to interrogate a potential target and identifies it as friendly or unknown. The BCIS system was designed to improve target identification performance and reduce fratricide (Doton, 1996). As with many such automated aids, however, it is not clear whether performance with the system is in fact significantly improved (Dzindolet, Pierce, Pomranky, Peterson, & Beck, 2001). Similarly, the Theater High Altitude Area Defense (THAAD) system is a complex weapons system designed to intercept enemy short and medium range ballistic missiles. After enemy information is input from the field soldier the THAAD system makes recommendations on engagement decisions. These automated recommendations are then presented to the soldiers for their approval. The result of following an errant or unreliable recommendation could have serious consequences on the outcome of a mission. Consequently, there is a pressing need for designing automation that supports military operators in command and control activities in ways that avoid such negative influences. In addition to the theoretical motivations for the present study, the experiments were also designed to provide information regarding these practical issues.

Hypotheses

Based on the review and the rationale outlined above the following hypotheses were made.

1. Automation will facilitate better human-system performance compared to manual conditions if it is reliable. This will include more correct and faster detections in the visual search studies. The automation will facilitate better human-system performance in the search and destroy mission.

2. The automation will show a greater effect on performance improvement under higher workload levels. In the target search task this performance improvement will become apparent as more distractors are added to the set size. In the search and destroy mission the automation will nullify the workload effect within stages as compared to the manual conditions where high workload will have a detrimental effect on performance.

3. Unreliable automation will cause a decrement in performance and that decrement will be greater for the automation in the decision-aiding stage as compared to the information analysis stage.

4. The decrement caused by unreliable automation hypothesized above will be different between the groups who receive different reliability levels. The decrement will be greater for those who have exposure to the lowest reliability level and increase linearly with higher reliability levels.

5. The Parasuraman et al. (2000) model will be scalable to other domains and to more complex task environments. The application of the model will provide information regarding primary and secondary task performance differences that would not be apparent without the use of the model. Further, the application of the model will reveal stage specific costs and benefits between manual and automation-aided conditions for both the primary and secondary tasks.

6. Human-system performance in the search and destroy task will be greater for the tasks in the current stage if the previous stage in the task was automated. This extends the first hypothesis to include a current performance gain from a previous exposure to automation.

EXPERIMENT ONE

Summary

A visual search paradigm was used to examine the effects of status information automation cueing in a target detection task. Manual and information automation conditions were manipulated with the size of the distractor set. Participants were required to respond to the presence or absence of a target in a time-limited trial. In the information automation condition, status information regarding target presence was presented to the participant. The participants were informed that the information automation was not perfectly reliable. A significant detection performance improvement was observed with the addition of the information automation. This improvement was more marked in the condition with the higher number of distractors. Additionally, detection performance declined when the information automation was invalid, without a corresponding increase in subjective measures of workload or confidence.

Introduction

The present study serves as the beginning of a series of planned comparisons between levels and stages of automation and the effect reliability levels have on action implementation. To date, studies looking for detection and/or performance differences by stage of automation have not utilized a common task environment. A visual search task was chosen as a simple simulation of a target identification environment, as used in BCIS or other automation systems for identification of friendly and enemy targets.

Methods

Participants

Four males and four females between the ages of 19 and 31 years (M = 20.00, SE = 2.65) served as paid participants. All participants were right-handed and reported normal or corrected-to-normal vision.

Experimental Design

A within-subjects design was employed in which two Automation Conditions (Manual, Information Automation) were combined factorially with Distractor Set Size (10, 20). Cue Validity was also manipulated within the Information Automation Condition.

Apparatus and Procedures

A visual search paradigm, in which participants were required to search a visual display for the presence or absence of a pre-defined target (╥) among similar distractors (╢, ╨, ╟, ╫), was employed. The display field emulated an artificial horizon consisting of 60% ground and 40% sky. Targets appeared in the ground portion of the display.

A trial began with the simultaneous display of the target and distractor elements, and lasted 2.5s, or until the participant responded. A target was present on 50% of the trials. Participants were required to respond, using the left or right-arrow keys, to the presence or absence of the target, respectively. Each participant completed three sessions of 100 trials in each automation condition. In half of the trials in each session, there were 10 distractors in the display; in the other half, 20 distractors were present. The trials were

randomized with respect to both the number of distractors and the presence or absence of a target. The order in which the sessions were presented was also randomized.

The Information Automation trials were identical to the Manual trials, with the exception that one of the elements in the display was highlighted, and participants were instructed that the highlighted element was likely to be the target if a target was present. On 67% of the trials, the highlighted element provided the participant with a valid cue to guide his/her response. On the remaining trials, this cue was inutile, e.g., a distractor element was highlighted when in fact there was a target present elsewhere in the field.

All participants achieved a 75% correct response criterion in practice trials in each automation condition (under the 10 distractor condition) before experimental data collection began. A modified NASA-TLX (Hart & Staveland, 1988) and subjective measures of confidence were administered after each 100 trials. The confidence measures were administered by asking each participant to provide a response to their ability to choose the correct course of action on a ten-point scale.

Results

Correct Responses

A correct response was defined as the outcome of a trial on which a participant either correctly detected the presence of a high-priority target – indicated by the initiation of a "fire" response – or correctly judged the absence of a high-priority target – specified by a "no fire" response. Mean percentages of correct responses were submitted to a 2×2 (Automation Condition × Set Size) repeated measures analysis of variance (ANOVA), revealing significant main effects of both Automation Condition, $F(1, 7) = 20.92$, $p < .05$, and Set Size, $F(1, 7) = 10.05$, $p < .05$, and an Automation Condition × Set Size interaction, $F(1, 7) = 7.19$, $p < .05$. For this and all analyses employed herein, (a) the Huynh-Feldt adjustment was applied, where appropriate, to guard against violations of the sphericity assumption; and (b) all post-hoc tests (pairwise t-tests) were corrected

using the Bonferroni procedure to maintain family-wise alpha-level at .05. It is apparent from Figure 3 that participants made fewer correct responses as the distractor set size increased from 10 to 20, and the number of correct responses was higher when the automated status information aid was employed.

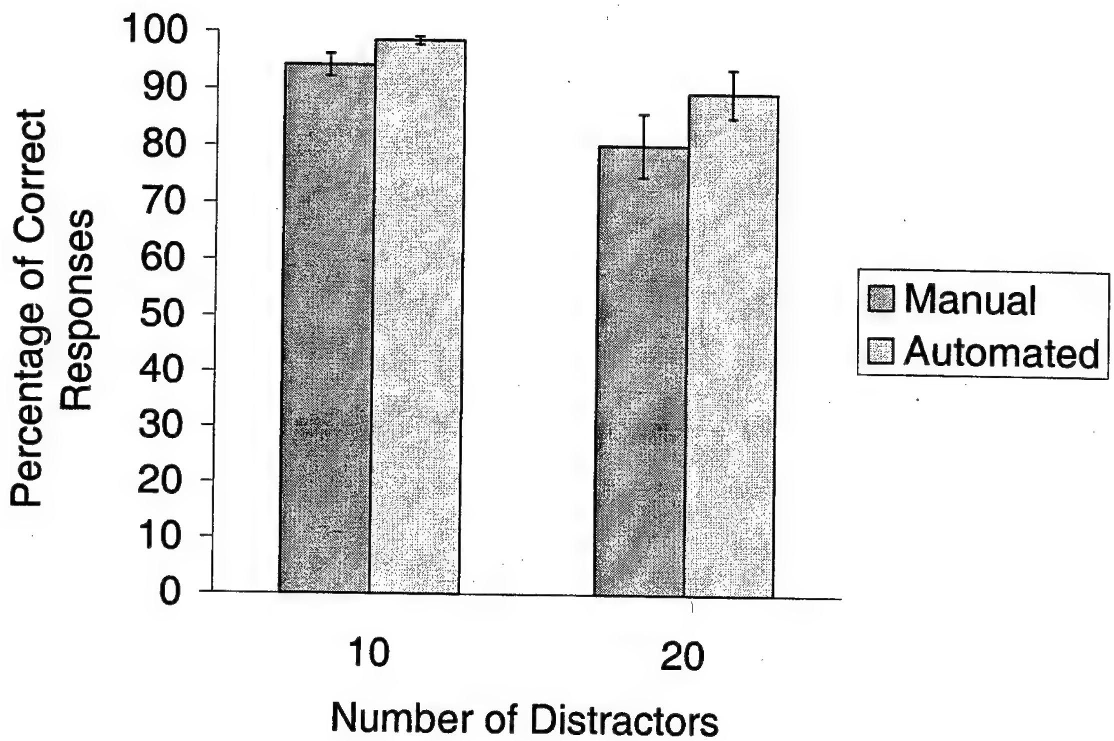


Figure 3. Mean percentage of correct responses as a function of automation condition and number of distractors. Error bars represent one standard error of the mean.

While the aforementioned analysis exposes the effect of automation on response accuracy, it fails to account for the validity of the automated cue. In order to examine these effects, mean percentages of correct responses were submitted to a 2 × 2 (Cue Validity × Set Size) repeated measures ANOVA. This analysis revealed significant main effects of Cue Validity, $\underline{F}(1, 7) = 26.29$, $\underline{p} < .05$, and Set Size, $\underline{F}(1, 7) = 11.96$, $\underline{p} < .05$, as

well as a Cue Validity × Set Size interaction, F(1, 7) = p < .05. This interaction, illustrated in Figure 4, suggests that cue validity does indeed mediate task performance, but only for sufficiently large set sizes.

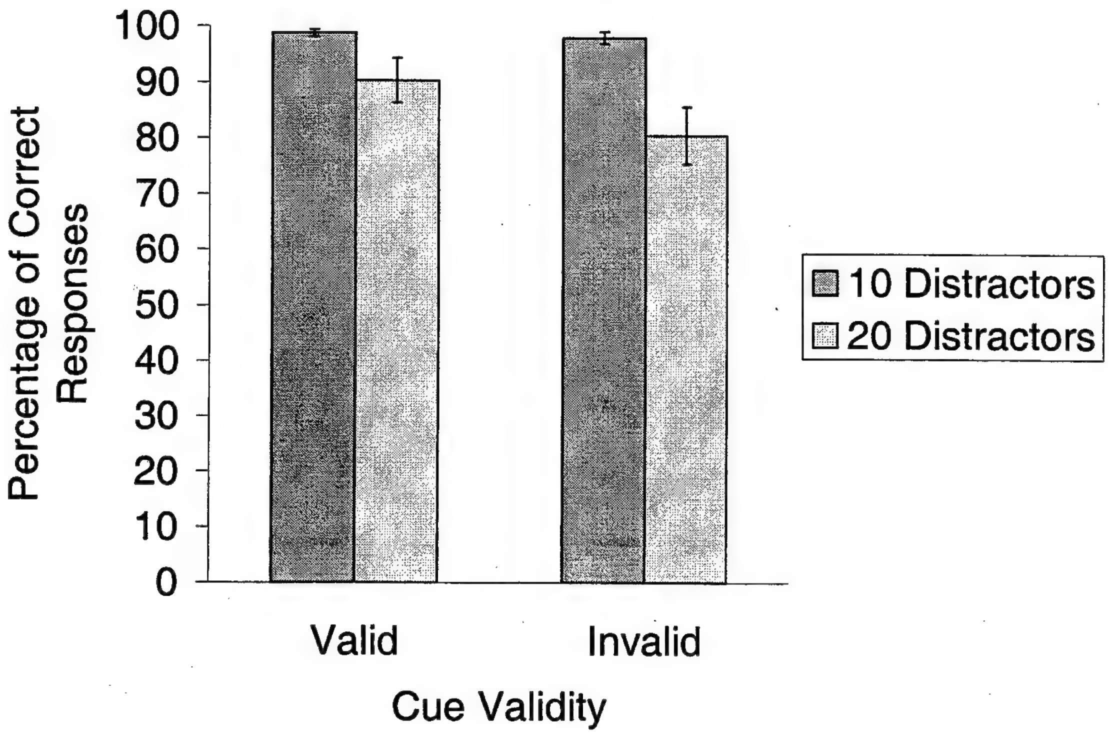


Figure 4. Mean percentage of correct responses as a function of cue validity and number of distractors. Error bars represent one standard error of the mean.

Response Time

Mean response times for correct responses were submitted to a pair of analyses analogous to those above. The Automation Condition × Set Size ANOVA disclosed significant main effects of both Automation Condition, $F(1, 7) = 16.65$, $p < .05$, and Set Size, $F(1, 7) = 7.0$, $p < .05$, and a significant interaction, $F(1, 7) = 19.61$, $p < .05$, illustrated in Figure 5. The response time data was similar to the percent correct data;

performance decreased as the number of distractors increased, and was better in the automated condition than in the manual condition, but only for the larger set size.

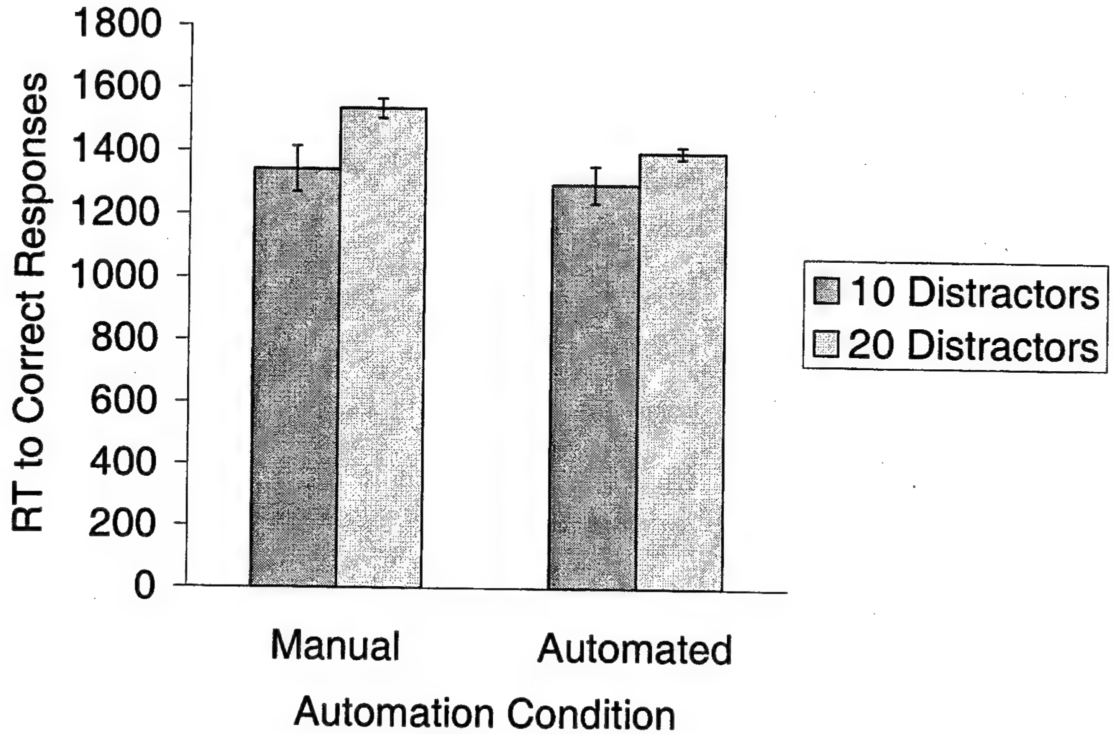


Figure 5. Mean response times (ms) for correct responses as a function of automation condition and number of distractors. Error bars represent one standard error of the mean.

The Cue Validity × Set Size ANOVA also displayed significant main effects of both independent variables (Cue Validity: $\underline{F}(1, 7) = 348.53$, $\underline{p} < .05$; Set Size: $\underline{F}(1, 7) = 9.83$, $\underline{p} < .05$) and a significant two-way interaction, $\underline{F}(1, 7) = 42.56$, $\underline{p} < .05$. As Figure 6 suggests, the source of this interaction is the nullification of the Set Size effect when the automated cue is valid. Response times increased with set size when invalid cues were employed. Furthermore, response times were higher in the invalid cueing condition.

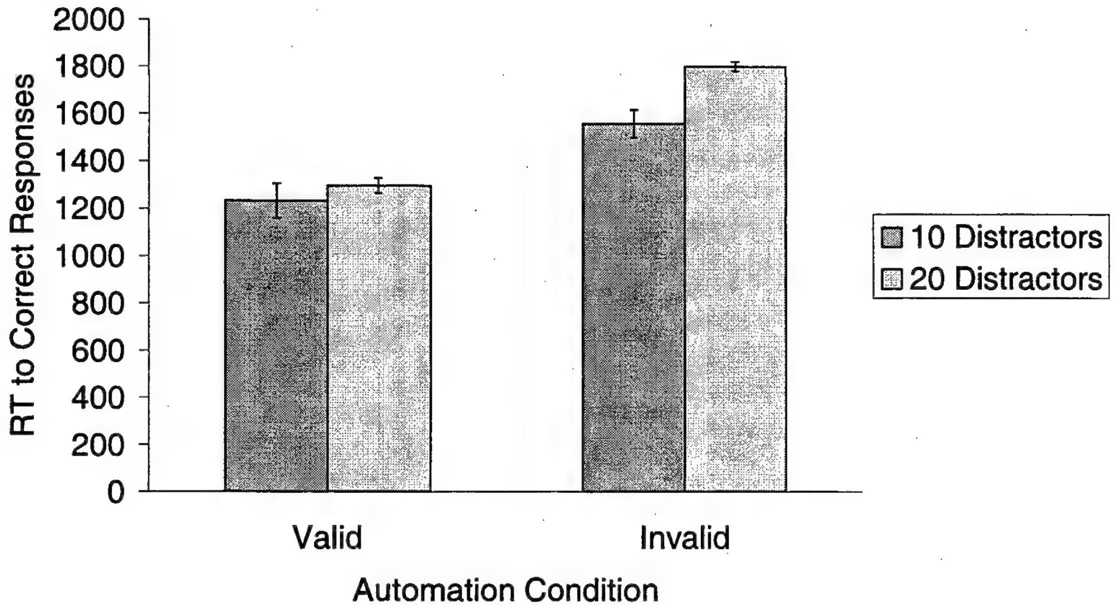


Figure 6. Mean response times (ms) for correct responses as a function of cue validity and number of distractors. Error bars represent one standard error of the mean.

Timeouts

A trial on which a participant failed to respond within the 2500ms response window was termed a "timeout." Mean percentages of timeouts were analyzed using analyses similar to those employed for the other dependent measures. The Automation Condition × Set Size analysis revealed significant main effects of both factors (Automation Condition: $\underline{F}(1, 7) = 6.84$, $\underline{p} < .05$; Set Size: $\underline{F}(1, 7) = 7.97$, $\underline{p} < .05$) and the interaction between them, $\underline{F}(1, 7) = 5.05$, $\underline{p} < .05$. As Figure 7 demonstrates, the tendency to time out was exacerbated by the addition of distractors to the set, as anticipated. Additionally, the presence of the automated cue reduced the number of timeouts. It should be noted that, by comparing these results with those displayed in Figure 3, one discovers that these timeouts account for the majority of the incorrect

responses. Mean percentages of timeouts were also analyzed in terms of cue validity, using procedures analogous to those employed for correct responses and response time, revealing a significant main effect of Set Size, $F(1, 7) = 5.84$, $p < .05$. All other sources of variance lacked significance.

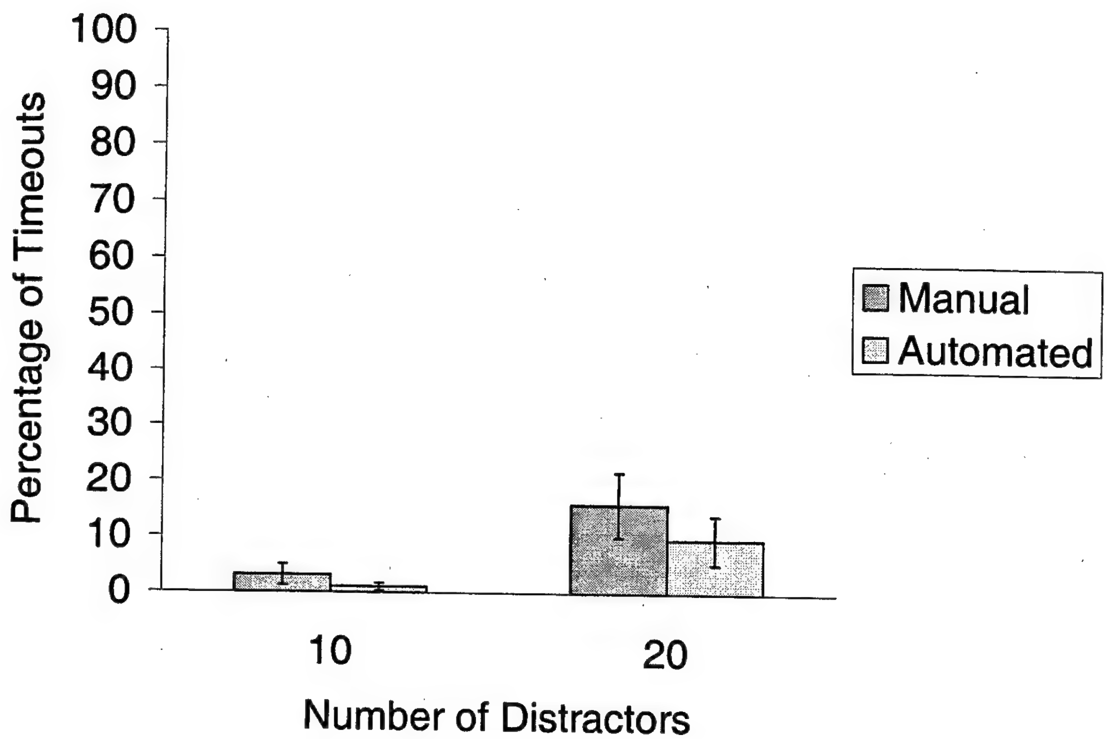


Figure 7. Percentage of timeouts as a function of automation condition and distractor set size. Error bars represent one standard error of the mean.

Confidence and Subjective Workload

Participants reported that they were equally confident in the manual ($M = 7.5$, $SE = 0.23$) and the automated status information ($M = 7.9$, $SE = 0.17$) conditions. They also reported similar ratings for the mental workload scales. Average TLX scores for the

manual ($\underline{M}$ = 32.19, $\underline{SE}$ = 4.23) and the automated ($\underline{M}$ = 28.98, $\underline{SE}$ = 3.98) conditions failed to show significant differences. .

Discussion

The results of this experiment indicated that a performance benefit was attained with the presence of the automated status information cue. Furthermore, this benefit increased with the addition of more distractors within the search field. These benefits were realized even though the participants were aware that the automated status information cue was not perfectly reliable. An examination of the participant's confidence level indicates that they were equally confident in their performance and reported similar subjective workload ratings regardless of the presence or absence of automation. In other words, the participants did not report that the automated status information cue was a hindrance to their performance.

These effects indicate that automated information cueing improved target identification performance under high target density conditions. The finding that the addition of the automation did not increase workload is important given that many forms of automation have been found to <u>increase</u> rather than decrease workload as they were designed to do (Wiener & Curry, 1980).

On the other hand, these results highlight the significant costs associated with the use of unreliable automation. The percentage of correct target identification responses was lower in the higher set size for the invalid cue, indicating that the participants were following the cue's recommendation, even though they were aware that the automation was not 100% reliable. This could represent a significant potential cost, given that perfect automation reliability cannot be assured.

In addition, these results demonstrate response time differences and an increased incidence of timeouts as distractor set size and cue validity were manipulated. The collective results suggest that an operator may have the most difficulty in the identification of friendly versus hostile entities when the battlespace is saturated, there is

limited time to respond, and the automation is less than 100% reliable. It is also apparent that the operator will perform significantly better under these circumstances if the automation is flawless. The trade-off in this circumstance is the price paid for an incorrect identification. The potential cost of following a strategy that adheres to the status cue made by the automation may not be apparent in a laboratory setting where the risk of an inaccurate identification and decision to fire has little consequence. When the potential cost is determined by the incidence of fratricide under battlespace management conditions, a high risk tasking environment, the cost of an inaccurate or untimely identification is deserving of serious consideration.

This study examined the effects of information automation on a visual search task, providing a benchmark for this task paradigm under these conditions. Subsequent studies will exploit this paradigm to investigate the effects on task performance of manipulations in the stage in which automation is present.

EXPERIMENT TWO

Summary

A visual search paradigm was used to examine the effects of information automation as well as decision-aiding automation in a target detection and processing task. Manual, information automation, and decision-aiding automation conditions were manipulated with the size of the distractor set. Participants were required to respond to the presence or absence of a target in a time-limited trial. In the information automation condition, status information regarding target presence was presented to the participant. The participants were informed that the information automation was not perfectly reliable. A significant improvement in detection performance was observed in the information automation condition. This improvement was more evident in the conditions with the higher number of distractors. Additionally, response times were improved when the information automation cue was present.

Introduction

This experiment serves as a continuation in a series of planned comparisons between levels and stages of automation and the effect reliability levels have on action implementation. The first study (Galster, Bolia, Parasuraman, & Roe, 2001) compared target detection in a manual and an automated information status cue in a basic visual search task. A visual search task was chosen as a simple simulation of a target identification environment, as used in BCIS or other automated systems for identification of friendly and enemy targets. The results indicated that there was a performance benefit attained with the presence of the automated aid, and that this benefit increased with the number of distractors. Moreover, these results were obtained without a concomitant increase in subjective workload. However, performance suffered when the automated cue was unreliable in the highest distractor set size, indicating an over-reliance on

automation, which is consistent with the results of Yeh et al. (1999), who found that target detection performance increased with valid cues but decreased with invalid cues.

To date, studies looking for detection and/or performance differences by stage of automation have not utilized a common task environment. The present study utilized the same basic visual search task as the previous study but included an automated decision-aiding condition in addition to the manual and information automation conditions. Thus, we can compare the results from the first study to the present study with more confidence than we could if different task environments were utilized.

Methods

Participants

Four males and four females between the ages of 18 and 28 years ($\underline{M}$ = 21.125, $\underline{SE}$ = 1.25) served as paid participants. All participants were right-handed and reported normal or corrected-to-normal vision.

Experimental Design

A 5 × 3 × 2 within-subjects design was employed in which 5 Automation Conditions (Manual (M), Information Automation (IA), Decision-Aiding (DA), Co-Located IA + DA, Separated IA + DA) were combined factorially with 3 Distractor Set Sizes (10, 20, 30). Cue Validity was also manipulated within each of the 15 treatments, to the extent that, in any given treatment, 70 percent of the cues were valid cues.

Apparatus and Procedures

A visual search paradigm, in which participants were required to search a visual display for the presence or absence of a pre-defined target (╥) among similar distractors

(╣, ╩, ╠, ╬), was employed. The display field emulated an artificial horizon consisting of 60% ground and 40% sky. Similar to the previous study, targets appeared only in the ground portion of the display.

All trials began with the presentation of a black fixation circle for 250ms at the center of the display, followed by an interval of 1s in which the display was blank except for the artificial horizon. This was followed by the presentation of the automation cue(s) on the artificial horizon that lasted 300ms. The IA cue (a red plus sign) was always located in the green target area while the DA cue ("fire" or "no fire") was located in the blue-sky portion unless co-located with the IA cue. The automated cue(s) were cleared for 500ms and the target and distractor items were presented for 2.5s, or until the participant initiated a response. Trials were separated by an inter-trial interval of 2s. A target was present on 50% of the trials. Participants were required to respond, using the left or right-arrow keys, to the presence or absence of the target, respectively. Each participant completed three sessions of 150 trials in each automation condition. There were an equal number of trials in each session representing each of the three distractor set sizes. The trials were randomized with respect to both the number of distractors and the presence or absence of a target. The order in which the sessions were presented was also counterbalanced.

All participants achieved a 75% correct response criterion in practice trials in each automation condition (under the 10 distractor condition) before experimental data collection began.

Results

Correct Responses

A correct response was defined as the outcome of a trial on which a participant correctly detected the presence of a high-priority target – indicated by the initiation of a "fire" response – or correctly judged the absence of a high-priority target – specified by a

"no fire" response. Mean percentages of correct responses were submitted to a 5 × 2 (Automation Condition × Set Size) ANOVA, revealing significant main effects of both Automation Condition, $F(4, 24) = 49.92$, $p < .05$, and Set Size, $F(2, 12) = 11.90$, $p < .05$, and an Automation × Set Size interaction, $F(8, 48) = 10.08$, $p < .05$.

As Figure 8 illustrates, the source of the Automation Condition × Set Size interaction is the lack of an effect of set size under the two combined conditions. In all other conditions, increases in set size occasioned decreases in the percentage of correct responses. Inspection of Figure 8 also reveals differences in search performance as a function of Automation Condition. Specifically, participants made more correct responses under the IA condition than under any of the other automation conditions or the manual control. Performance under the DA automation condition was not different from performance under the manual condition for any set size. Surprisingly, participants made significantly fewer correct responses under the combined automation conditions, performance under which was always below or indistinguishable from performance under the other conditions.

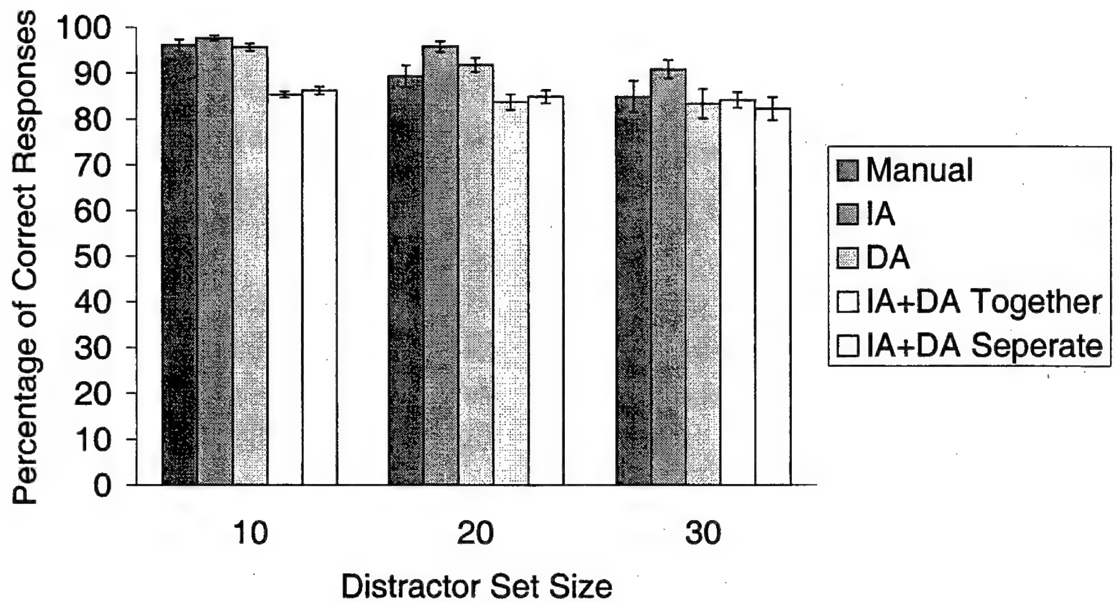


Figure 8. Mean percentages of correct responses as a function of automation condition and set size. Error bars represent one standard error of the mean.

While the aforementioned analysis exposes the effect of different automation schemes on correctness of response, at least one source of variance remains unaccounted for: namely, the differences in the percentages of correct responses that might occur within the four automated conditions as a function of cue validity. In order to examine these effects, mean percentages of correct responses were submitted to a 4 × 3 × 2 (Automation Condition × Set Size × Cue Validity) repeated measures ANOVA. This analysis revealed significant main effects of Set Size, $F(2, 12) = 11.14$, $p < .05$, and Cue Validity, $F(1, 6) = 6.80$, $p < .05$. Each of the two-way interactions was also significant (Automation Condition × Set Size, $F(6, 36) = 3.50$; Automation Condition × Cue Validity, $F(3, 18) = 38.35$; Set Size × Cue Validity, $F(2, 12) = 6.71$), all $p < .05$. All other sources of variance lacked statistical significance ($p > .05$).

Response Times

Mean response times for correct responses were analyzed using similar procedures. The 5 × 3 (Automation Condition × Set Size) ANOVA disclosed significant main effects of Automation Condition, $F(4, 24) = 38.59$, $p < .05$, and Set Size, $F(2, 12) = 95.58$, $p < .05$ and their interaction, $F(8, 48) = 6.07$, $p < .05$. In Figure 9, in which mean response times are plotted as a function of automation condition and set size, it is apparent that the IA and combined automation conditions engendered significantly lower response times than either the DA or Manual conditions.

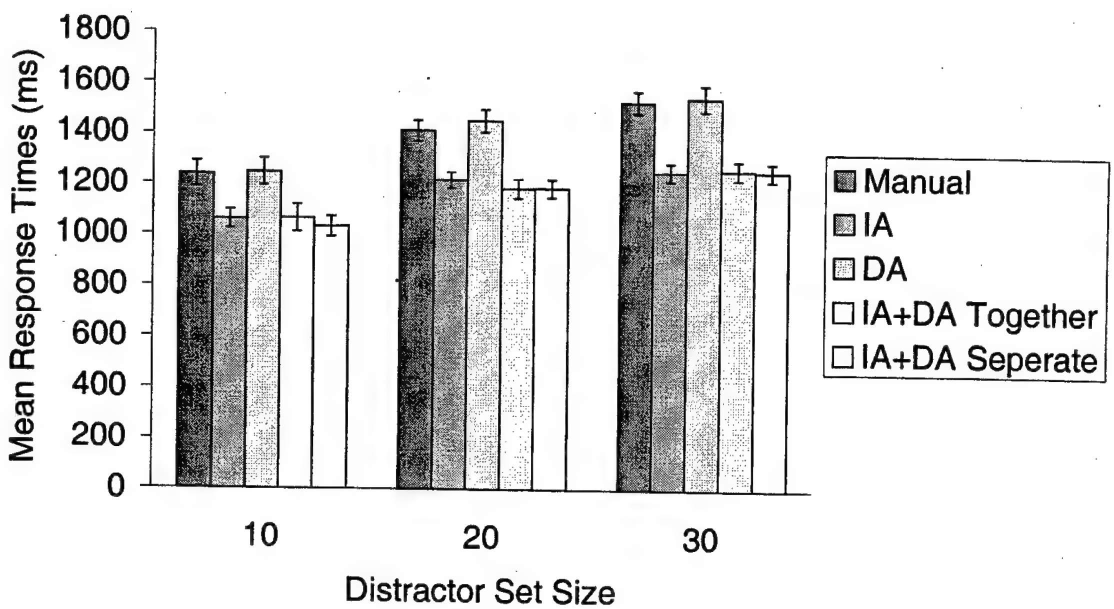


Figure 9. Mean response times (ms) to correct responses as a function of automation condition and set size. Error bars represent one standard error of the mean.

There was also a linear increase in response times associated with an increase in set size ($p < .05$), as expected in a search task involving conjunctions of features (Treisman & Gelade, 1980). As is seen in Table 2, the slopes of the response time vs. set size function, which can be interpreted as the temporal cost associated with the addition

of each distractor element to the set, were greater under the manual and DA automation conditions than under the other three conditions. This, along with the performance data described above, suggests that participants did not make use of the automated recommendation, but instead performed a serial search of the display set, as in the manual condition.

Automation Condition	Slope	t	p
Manual	15ms/distractor	4.67	< .05
IA	10ms/distractor	3.82	< .05
DA	15ms/distractor	4.27	< .05
IA + DA Together	10ms/distractor	3.24	< .05
IA + DA Separate	11ms/distractor	4.23	< .05

Table 2. Results of simple linear regression analyses on response time as a function of set size under each of the automation conditions.

The results of the 4 × 3 × 2 (Automation Condition × Set Size × Cue Validity) ANOVA are displayed in Table 3. As is apparent from the table, all main effects and interactions were statistically significant ($p < .05$).

Factors	df	F	p
Automation Condition (A)	(3, 18)	18.03	< .05
Set Size (B)	(2, 12)	159.89	< .05
Cue Validity (C)	(1, 6)	295.94	< .05
A × B	(3, 36)	2.63	< .05
A × C	(3, 18)	119.93	< .05
B × C	(2, 12)	22.71	< .05
A × B × C	(3, 36)	6.46	< .05

Table 3. Results of the 4 × 3 × 2 (Automation Condition × Set Size × Cue Validity) repeated measures ANOVA to which the response time data were subjected.

On average, participants took 400-650ms longer to locate a high-priority target when the cue was invalid. On the other hand, the performance data indicated that participants performed better under these conditions with invalid cues than they did with valid cues, implicating a possible speed-accuracy tradeoff.

Incorrect Responses

While the analyses conducted above are informative as to the speed and accuracy with which observers acquired targets under certain combinations of set size and automation condition, they fail to explain how errors were committed. In this experiment, there were two types of incorrect responses: those that were incorrect in the sense that the subject initiated an inappropriate response, and those that were incorrect in the sense that the participant failed to respond within the 2500ms response window. For the purposes of this chapter, the former will be termed "incorrect responses," and the latter "timeouts." Table 4 presents the mean percentages of correct responses, incorrect responses, and timeouts as a function of automation condition and set size.

It is clear from Table 4 that the data from the manual and DA automation conditions are very similar, as was the case for other dependent measures described. In both cases, the percentage of timeouts was very low when there were only 10 distractors present (i.e., when the search task was comparatively easy), and increased steadily as more distractors were added to the set. Indeed, in the 30-distractor condition, participants failed to initiate a response nearly 10% of the time. Under the 10- and 20-distractor conditions, participants had more incorrect responses than timeouts, but as the number of distractors increased to 30, the number of timeouts surpassed the number of incorrect responses.

Automation Condition	Set Size		
	10	20	30
Manual			
Percent Correct	96.09	89.33	84.86
Percent Incorrect	2.95	5.81	6.09
Percent Timeouts	0.95	4.86	9.05
IA			
Percent Correct	97.70	95.79	90.83
Percent Incorrect	2.01	2.10	3.82
Percent Timeouts	0.29	2.10	5.36
DA			
Percent Correct	95.70	91.78	83.30
Percent Incorrect	3.54	4.59	6.77
Percent Timeouts	0.76	3.63	9.93
IA + DA Together			
Percent Correct	85.35	83.68	84.10
Percent Incorrect	14.07	13.84	11.58
Percent Timeouts	0.57	2.48	4.32
IA + DA Separate			
Percent Correct	86.22	84.84	82.20
Percent Incorrect	13.21	13.90	13.52
Percent Timeouts	0.57	1.26	4.28

Table 4. Percentages of correct responses, incorrect response, and timeouts as a function of automation condition and set size.

Under the IA automation condition, the pattern of timeouts was similar to that obtained in the manual and DA conditions, but on a smaller scale, with participants failing to respond about half as often. Additionally, there is little variance in the percentage of incorrect responses as a function of set size. The data for the combined conditions were similar to those for the IA condition with respect to timeouts, but much elevated in terms of the percentage of incorrect responses. On average, participants

responded incorrectly about 13% of the time when one of the combination automation schemes was employed.

Discussion

The results of this experiment indicate that a performance benefit was achieved with the presence of the IA status cue. Furthermore, this benefit increased with the addition of more distractors within the search field. These benefits were realized even though the participants were aware that the automated IA cue was not perfectly reliable.

These effects indicate that automated information cueing improved target identification performance under high target density conditions. Thus the benefit of real-world battlefield or air defense identification systems might best be realized in complex, dense engagements, when the operator is likely to be already near their peak level of workload.

In addition, these results demonstrated response time gains with the presence of the IA cue, by itself or in conjunction with the DA cue, indicating that location is the enduring variable in reducing response times in this task.

EXPERIMENT THREE

Summary

The visual search paradigm was used to examine the effects of information automation and decision-aiding automation in a target detection and processing task. Manual, information automation, and decision-aiding automation conditions were manipulated with the size of the distractor set. Participants were required to respond to the presence or absence of a target in a time-limited trial. Reliability level (90%, 70%, 50%) of the automation was manipulated as a between-subjects variable. Each reliability level group was comprised of eight volunteers for a total of 24 participants. Results indicated that the information automation cue condition engendered an increase in correct responses and a reduction in search times, regardless of set size or automation reliability level. On the other hand, the presence of a decision-aiding cue differentially affected performance on all dependent measures as a function of both set size and automation reliability, alone or in concert with an information automation cue.

Introduction

The present study again served as a continuation of a series of planned comparisons between levels and stages of automation and the effect reliability levels have on action implementation. The first study (Galster et al., 2001) compared target detection in a manual and an automated information status cue in a basic visual search task. The results indicated that there was a performance benefit attained with the presence of the automated aid, and that this benefit increased with the number of distractors. Moreover, these results were obtained without a concomitant increase in subjective workload. However, performance suffered when the automated cue was unreliable in the highest distractor set size, indicating an over-reliance on automation.

The second study (Galster, Bolia, & Parasuraman, 2002), based on the same target identification task, revealed that a similar performance benefit was achieved with the

presence of the information automation status cue indicating only the location of the target. Furthermore, this benefit increased with additional distractors within the search field. These benefits were realized even though the participants were aware that the information automation cue was not perfectly reliable. The second study also contained a decision-aiding automation cue that suggested a possible action to the participant. This cue did not produce a performance benefit however, over and above the manual un-aided condition except when it was combined with the information automation cue.

These effects indicate that automated information cueing improves target identification performance under high target density conditions. In addition, the results demonstrated response time gains with the presence of the information automation (IA) cue, by itself or in conjunction with the decision-aiding (DA) cue, indicating that location is the enduring variable in reducing response times and increasing correct detections in this task.

To date, studies looking for detection and/or performance differences by stage of automation have not utilized a common task environment. The present study utilized the same basic visual search task with a manual, automation information, and decision-aiding cueing. But, unlike the previous study this study examined the manipulation of the reliability level of the automation as a between-subjects variable. Thus again, comparisons of the results from the first two studies to the present study can be made with more confidence than if a different task environments were utilized.

Methods

Participants

Fourteen males and ten females between the ages of 18 and 32 years (M = 21.92, SE = 2.35) served as paid participants. All participants were right-handed and reported normal or corrected-to-normal vision.

Experimental Design

A mixed design was employed in which 4 Automation Conditions (Manual (M), Information Automation (IA), Decision-Aiding (DA), Co-Located (IA + DA)) were combined factorially with 3 Distractor Set Sizes (10, 20, 30) to serve as within-subjects variables. Automation Reliability (90%, 70%, 50%) was manipulated as a between-subjects variable.

Apparatus and Procedures

A visual search paradigm, in which participants were required to search a visual display for the presence or absence of a pre-defined target (╦) among similar distractors (╣, ╩, ╠, ╬), was employed. The display field emulated an artificial horizon consisting of 60% ground and 40% sky. Targets appeared only in the ground portion of the display.

All trials began with the presentation of a black fixation circle for 250ms at the center of the display, followed by an interval of 1s in which the display was blank except for the artificial horizon. This was followed by the presentation of the automation cue(s) on the artificial horizon that lasted 300ms. The IA cue (a red plus sign) was always located in the green target area while the DA cue ("fire" or "no fire") was located in the blue-sky portion unless co-located with the IA cue. The automated cue(s) were cleared for 500ms and the target and distractor items were presented for 2.5s, or until the participant initiated a response. Trials were separated by an inter-trial interval of 2s. A target was present on 50% of the trials. Participants were required to respond, using the left or right-arrow keys, to the presence or absence of the target, respectively. Each participant completed eight sessions of 150 trials during data collection. There were an equal number of trials in each session representing each of the three distractor set sizes. The trials were randomized with respect to both the number of distractors and the presence or absence of a target. The sessions consisted of two manual conditions, one of each IA, DA, IA+DA conditions of the prescribed reliability levels and one of each IA, DA, IA + DA conditions where the automation was perfectly reliable. The conditions

were counterbalanced with respect to the automation condition and the order of the level of reliability.

All participants achieved a 75% correct response criterion in practice trials in each automation condition (under the 10 distractor condition) before experimental data collection began.

Results

Correct Responses

A correct response was defined as the outcome of a trial on which a participant either correctly detected the presence of a high-priority target – indicated by the initiation of a "fire" response – or correctly judged the absence of a high-priority target – specified by a "no fire" response. Mean percentages of correct responses were submitted to a $4 \times 3 \times 3$ (Automation Condition × Set Size × Reliability Level) ANOVA. The analysis revealed significant main effects of both Automation Condition, $F(3, 63) = 30.78$, $p < .05$, and Set Size, $F(2, 42) = 59.74$, $p < .05$, an Automation Condition × Set Size interaction, $F(6, 126) = 15.38$, $p < .05$, and an Automation Condition × Automation Reliability interaction, $F(6, 63) = 15.14$, $p < .05$. Neither of the other interactions was a significant source of variance. The Automation Condition × Set Size and Automation Condition × Automation Reliability interactions are depicted in Figures 10 and 11, respectively.

Inspection of Figure 10 reveals differences in search performance as a function of automation condition. Specifically, participants made more correct responses under the IA condition than under any of the other automation conditions or the manual control. Performance under the DA automation condition was only marginally different from performance under the manual condition for any set size. Further, correct response performance appears relatively stagnant under the combined IA + DA Automation condition across the three set sizes.

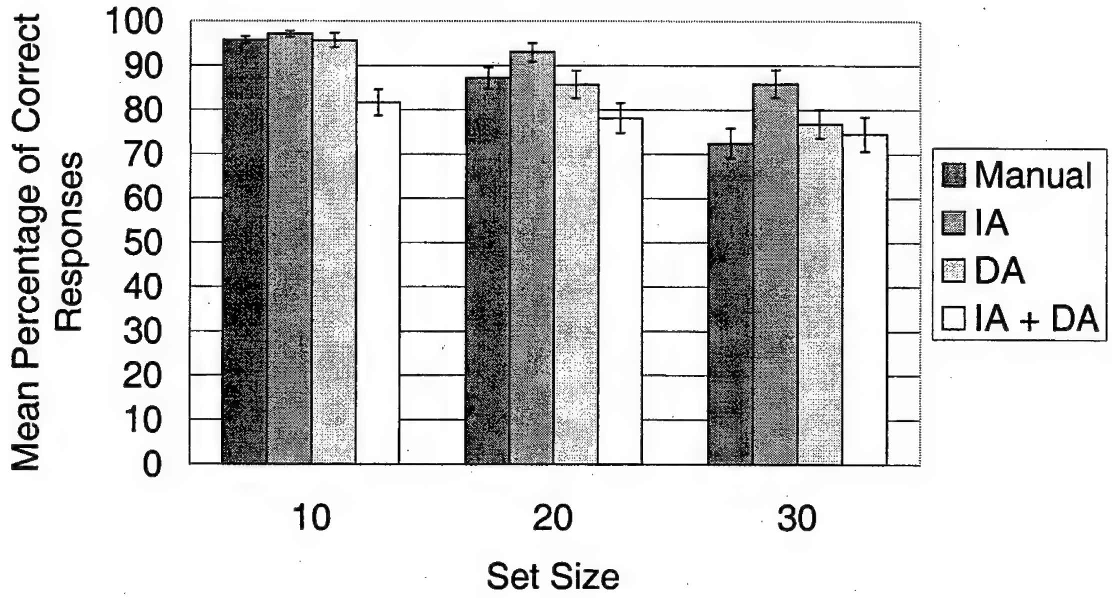


Figure 10. Mean percentages of correct responses as a function of automation condition and set size. Error bars represent one standard error of the mean.

Figure 11 illustrates the interaction between automation reliability and automation condition. A major source of this interaction appears to be the increase in correct responses associated with increased automation reliability. Figure 11 also suggests that the combination of information automation and decision aiding cues can lead to automation-induced complacency, a reduction in decision accuracy that can be caused by over-reliance in the automation. This effect is especially evident in the 50% and 70% reliability conditions.

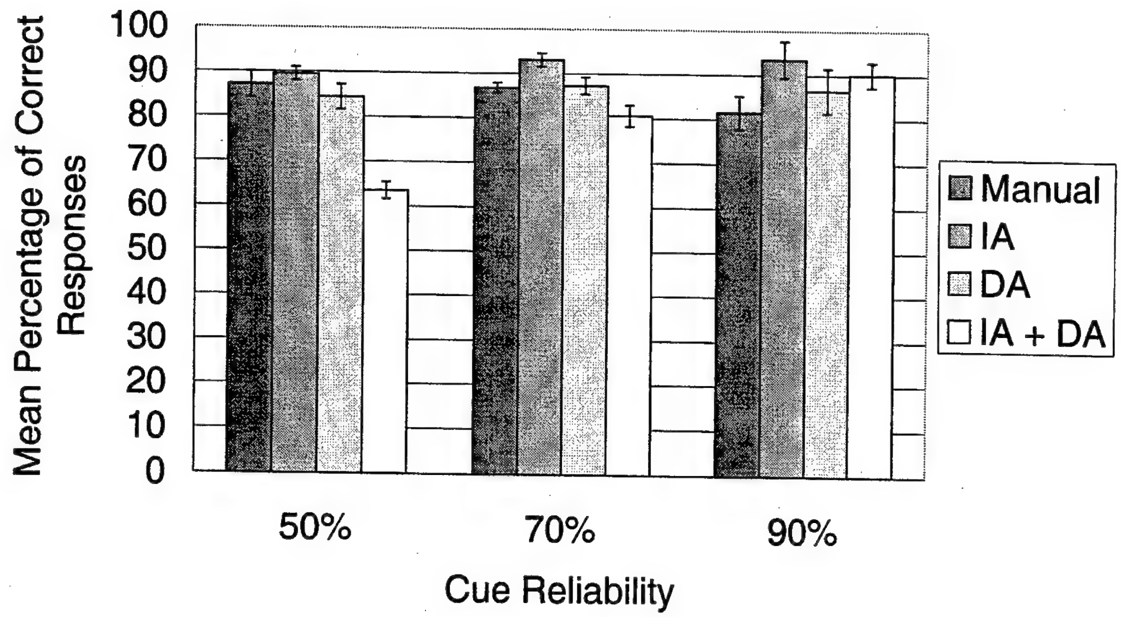


Figure 11. Mean percentages of correct responses by automation condition as a function of automation reliability level. Error bars represent one standard error of the mean.

Response Times

Mean search times of correct responses were submitted to an ANOVA analogous to that conducted for the percentages of correct responses. This analysis revealed significant main effects of both Automation Condition, $F(3, 63) = 37.85$, $p < .05$, and Set Size, $F(2, 42) = 85.55$, $p < .05$, an Automation Condition × Set Size interaction, $F(6, 126) = 5.32$, $p < .05$, and an Automation Condition × Automation Reliability interaction, $F(6, 63) = 8.22$, $p < .05$. None of the other sources of variances was significant ($p > .05$). The Automation Condition × Set Size and Automation Condition × Automation Reliability interactions are presented in Figures 12 and 13, respectively.

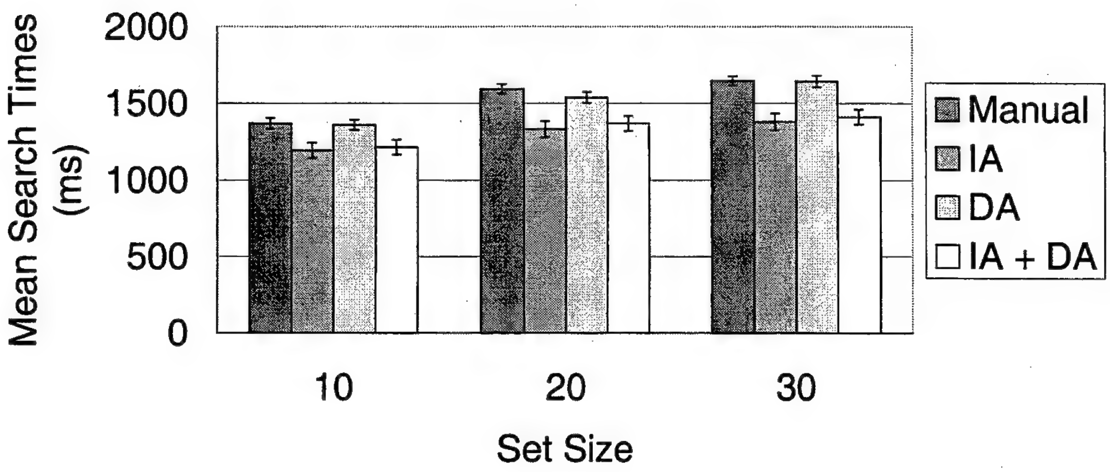


Figure 12. Mean search time (ms) as a function of set size and automation condition. Error bars represent one standard error of the mean.

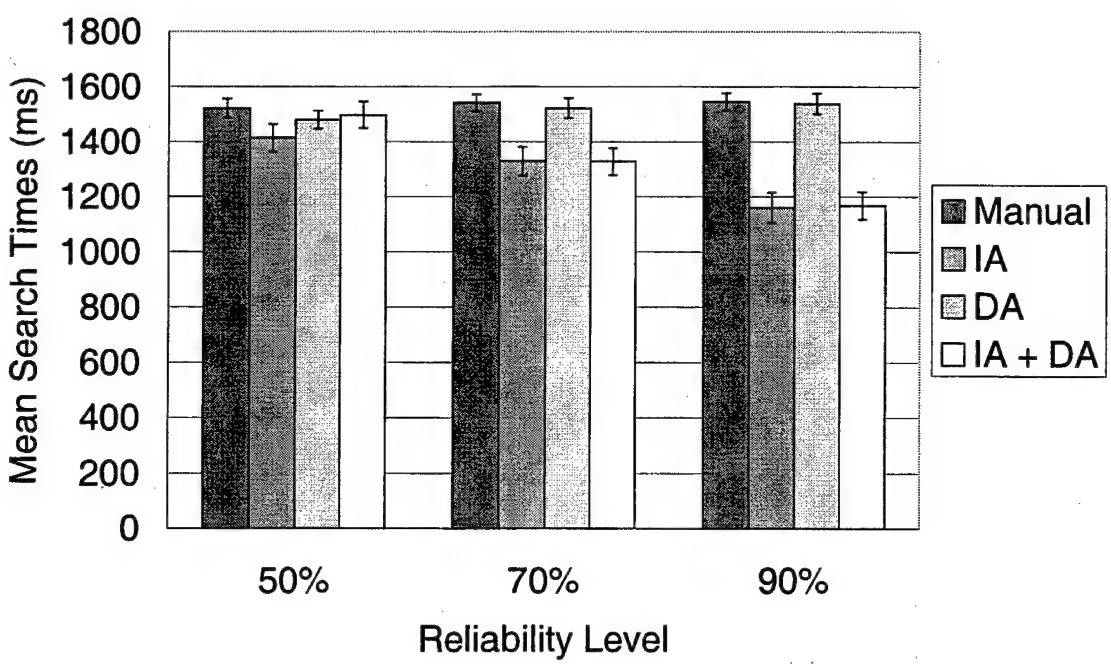


Figure 13. Mean search time (ms) as a function of automation reliability and automation condition. Error bars represent one standard error of the mean.

The results depicted in Figure 13 are consistent with those obtained in previous studies in this series (Galster et al. 2001; Galster et al. 2002). Namely, search times were reduced when an IA cue was present either alone or in conjunction with a DA cue. This effect is exacerbated under higher set sizes. This IA dominance effect is also visible in Figure 13, which demonstrates a decrease in search time with increasing automation reliability. This suggests that target acquisition and action implementation in a saturated complex visual field is enhanced most effectively by the presence of a reliable IA cue providing location information.

Timeouts

Another source of variance not accounted for in the examination of correct responses is the number of trials in which a response is not made in the prescribed 2500ms allowed, termed a timeout. Mean percentages of total trials that resulted timeouts were submitted to an ANOVA analogous to that conducted for the previous two analyses. This analysis revealed significant main effects of both Automation Condition, $\underline{F}(3, 63) = 13.16$, $\underline{p} < .05$, and Set Size, $\underline{F}(2, 42) = 41.35$, $\underline{p} < .05$, an Automation Condition × Set Size interaction, $\underline{F}(6, 126) = 11.36$, $\underline{p} < .05$, and an Automation Condition × Automation Reliability interaction, $\underline{F}(6, 63) = 2.28$, $\underline{p} < .05$. None of the other sources of variances was significant. The Automation Condition × Set Size and interaction is presented in Figure 14. Timeouts clearly increase with increases in set size, as expected. Of particular interest is the reduction of timeouts when the information automation cue was presented alone or combined with the decision-automation cue as compared with the other two automation conditions. Thus, when the IA cue was present, it not only reduced the search times, it also reduced the number of trials that ended due to a lack of a response.

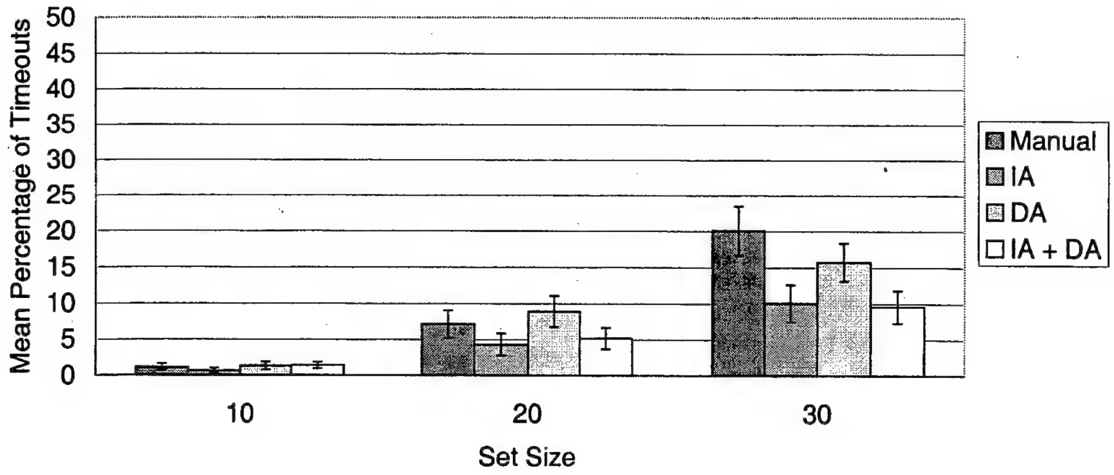


Figure 14. Percentage of timeouts as a function of set size and automation condition. Error bars represent one standard error of the mean.

Discussion

The results of this experiment indicate that a performance benefit was achieved with the presence of the IA status cue. Furthermore, this benefit increased with the addition of more distractors within the search field and persisted under all reliability levels. These benefits were realized even though the participants were aware that the automated IA cue was not perfectly reliable.

The results also revealed that a performance decrement was present in the IA+DA condition for both set size and differences in the reliability of the automation. Of particular interest is the decrement found with this condition under the 50% reliability rate. This is most likely due to over-reliance on the automation to give the correct guidance resulting in an automation induced complacency effect under those conditions.

In addition, these results demonstrate response time gains with the presence of the IA cue, by itself or in conjunction with the DA cue, indicating again that location is the enduring variable in reducing response times in this task. This effect was most prominent in the 90% reliability group but decreased as the reliability rate decreased.

EXPERIMENT FOUR

Introduction

The fourth study in this effort was designed to examine (a) the scalability of the Parasuraman et al. (2000) model, (b) the inclusion of all possible manual and automation control combinations, (c) the effect of workload differences, and (d) the effect of adding secondary task requirements to the operator. After the results of the visual search studies had been analyzed, an additional objective was added for the fourth study. The results of the previous studies strongly indicated that the information automation cue that provided the probable location of the target consistently led to a higher percentage and faster correct detections by the participants. This result was obtained despite set size differences and the reliability level of the cue, if all other cues had similar reliability levels.

The basic visual search task and the air-to-ground search and destroy task had similarities and differences. The differences include the complexity of the task, the duration of the task, and the addition of a secondary task. The similarities include the temporal compression, varying workload levels, and the stated objective of the tasks, which was to identify and respond to targets. Although the duration of the search and destroy mission was much longer than the visual search task, the temporal compression was similar in that the participants needed to complete a greater number of tasks in a relatively short period of time. The objective of each task was to identify a target (or group of targets) and initiate a response to that potential threat. In each case the target was clearly identified, or in the search and destroy mission, the targets were clearly prioritized according to their lethality potential. Thus, once the identification objective was met the decision of what action to take became obvious.

Due to the emphasis on the identification stage of the task (information automation in the visual search task and information acquisition in the search and destroy task) it was reasoned that overall task performance would be influenced by the early stage of the task, primarily under automated conditions. The results of the visual search task

support this reasoning. Performance benefits were observed when the information automation was present compared to when that stage was performed manually. To evaluate this potential effect in the search and destroy mission, another hypothesis was generated. Overall human-system performance would be greater if the information acquisition stage of the task was automated compared to when that stage was performed manually. In order to test this hypothesis the overall task measures were statistically analyzed by comparing them to the automation level in the information acquisition stage. The overall workload level was also included in the analyses as it was an overall task manipulation. The automation levels of the subsequent stages were not included in the analyses thus isolating any overall difference to the level of automation in the information acquisition stage and the level of workload for the overall task.

Methods

Participants

Eight male military pilots between the ages of 36 and 48 years (M = 39.75, SE = 1.39) served as volunteer participants. All pilots had extensive military fixed-wing flight experience ranging from 2000-3503 hours (M = 2714.83 hr, SE = 192.11) and reported normal or corrected-to-normal vision.

SIRE Facility

The Synthesized Immersion Research Environment (SIRE) facility (see Figure 15) contains a fixed-base cockpit situated in the center of a 40 foot diameter dome that includes a high-resolution, large field-of-view (70° vertical by 150° horizontal) interactive visual display. The cockpit for this experiment included three heads-down displays (HDD) and a single heads-up display (HUD). Control and task input actions from the pilots were achieved utilizing a sidestick and throttle, both containing switches with multiple functions. The SIRE facility is a conglomeration of several individual

systems that function as a system of systems. The major components in SIRE include the cockpit, the image generation hardware and software, the terrain database, the virtual battlespace software, and the software that controls and interacts with all of these systems. The basic functions of these systems will be described in the paragraphs below.

Cockpit

The cockpit in the SIRE facility is generic but resembles a F-16 cockpit due to the side-stick configuration. The cockpit is fixed with regard to roll, pitch, and yaw axes. The cockpit sits on a hydraulic platform that can be raised and lowered. The platform, in this experiment, was raised to bring the pilot's eye to the height of the design center of the visual dome (7' 7'').

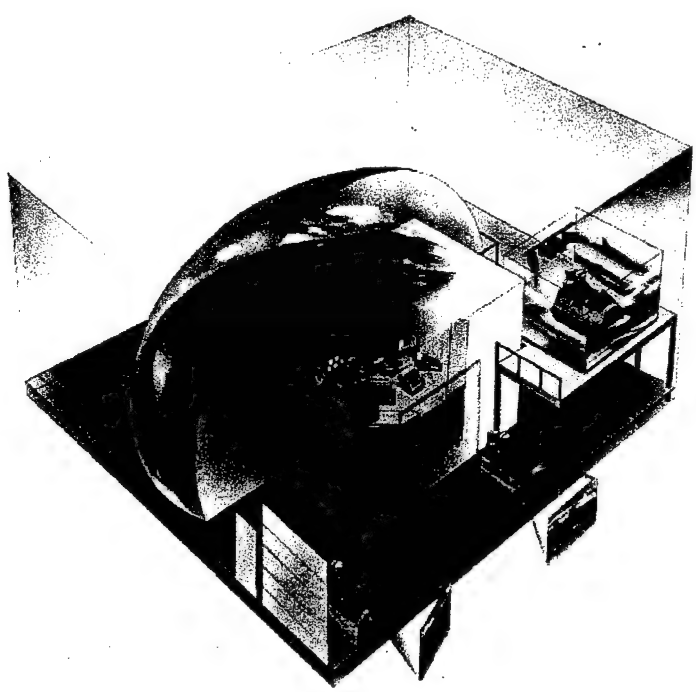


Figure 15. The SIRE facility.

Specific cockpit details are included in Appendix A. These include the display configuration used for this experiment and the functions of all relevant control input switches. Figure 16 illustrates the general layout of the cockpit HDD displays and the HUD.

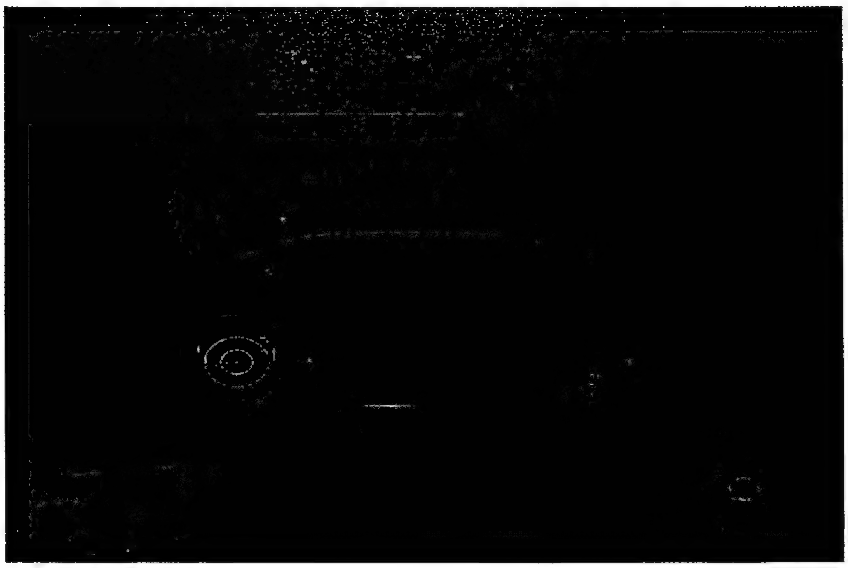


Figure 16. Reproduction of the cockpit as seen from the pilot's point of view.

Image Generation

The viewing screen comprises a 150° by 70° section of the forty-foot diameter dome with a matte white, unity gain surface coating. The visual screen displays the image a pilot would see extending 50° to -20° vertically and –75° to 75° horizontally with respect to the design center of the dome (x, y, z = 0°, 0°, 0°). The system incorporates distortion corrections to map the logical display plane perspective geometry from the Image Generator (IG) onto the physical screen surface. The IG positions the logical display planes with respect to the dome design eyepoint according to the criteria

in the Figure 17. Channels 1 and 3 were turned off during this experiment leaving four channels operational.

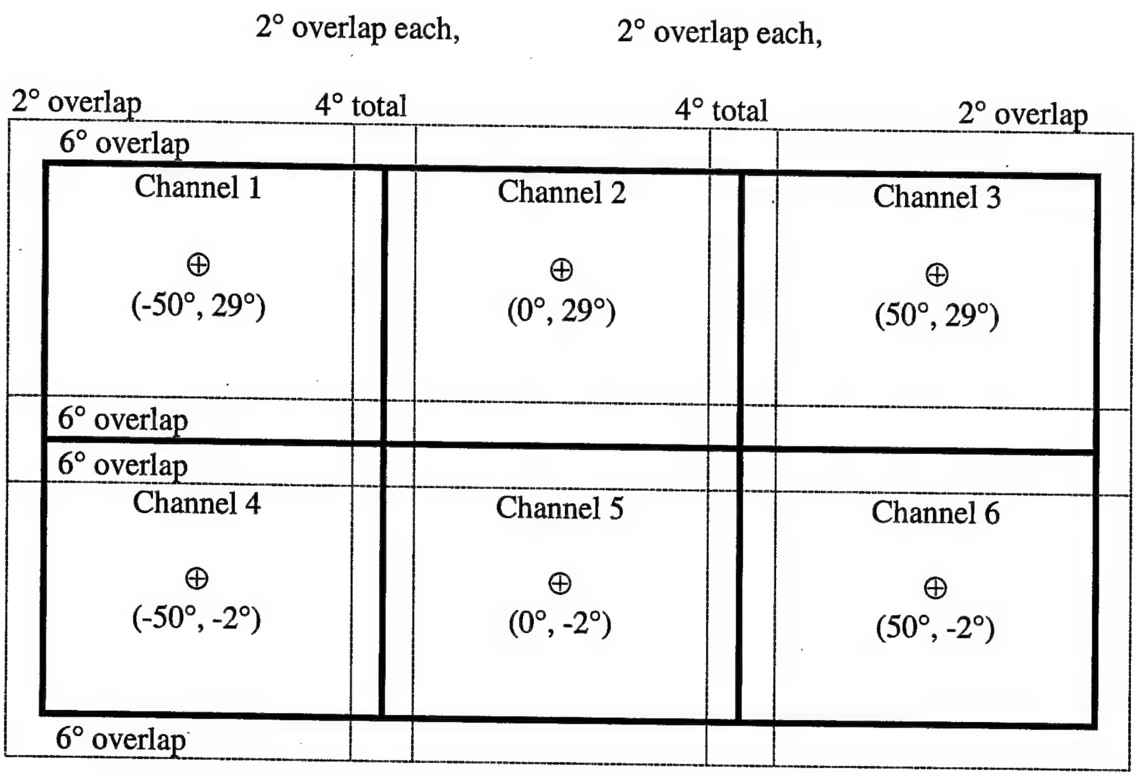


Figure 17. Channel configuration for viewing planes.

A dedicated dual 350Mhz processor computer controlled each video channel. The video signal from each channel was sent to a projector where it was projected at a high resolution (1280 × 1024) onto the screen.

Terrain Database

The terrain database extends from 31° to 33° North latitude and 109° to 110° West longitude. This area is centered approximately 60nm east of Tucson, AZ, USA.

Virtual Battlespace Environment

The virtual battlespace environment was created with the ModSAF (Modular Semi-Automated Forces) software package version 5.0. It allowed a single operator to create and control computer generated forces and save these configurations to a file for future use. ModSAF was used in this experiment to model the entities that were present for each of the missions flown. For additional information, Middendorf, Galster, and Brown (2003) describe the utilization of ModSAF in the present experiment.

Control Software

An overview of the software architecture used in the present study is illustrated in Figure 18. This illustration reflects the functional distribution of the computer programming models. All of the simulation software models communicated with each other using Microsoft's distributed communication tool (DCom). The distributed interactive simulation (DIS) server translated between DCom packets and protocol data units (PDUs), which conform to the IEEE DIS protocol. All communications took place on a local area network using 100 Mb/sec Ethernet cables.

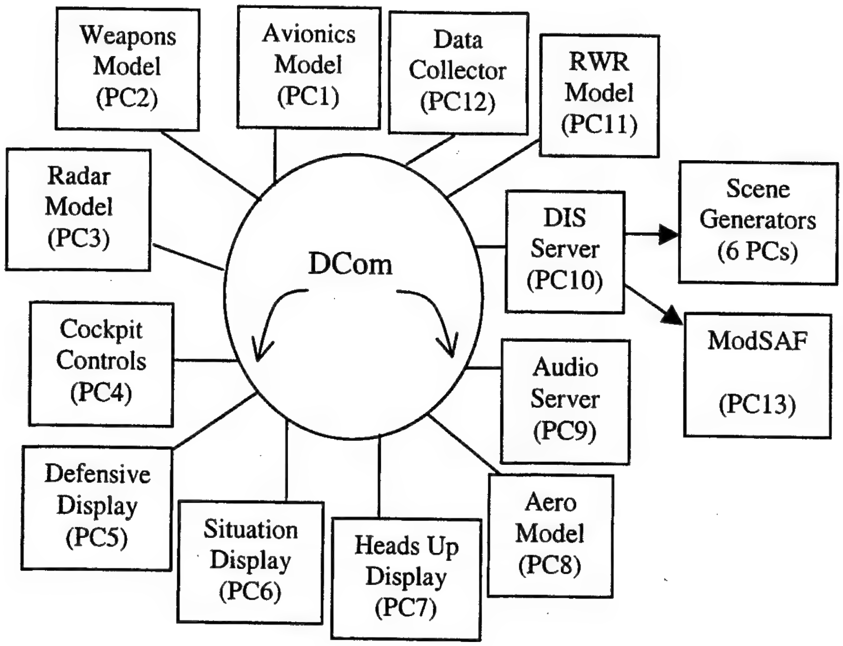
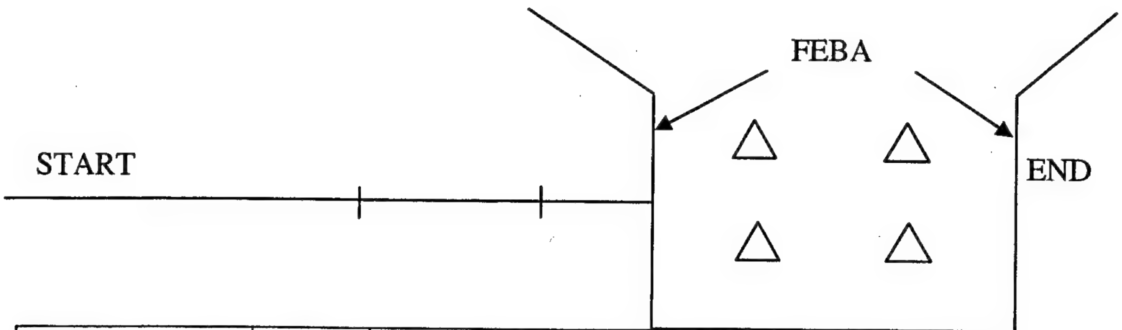


Figure 18. Software model configuration.

Overview of Flight Task

For each mission scenario, pilots were instructed to follow a waypoint driven flight plan at prescribed altitudes and airspeeds. The target engagement area was denoted by a forward edge of the battle area (FEBA) line depicted on one of the HDDs. The waypoints segmented the mission into four functional stages, which mapped on to the Parasuraman et al. (2000) model. Waypoint placement, associated stage transition points, distance to the FEBA, and assigned altitudes and airspeeds are illustrated in Figure 19. Specific details of the mission requirements, control input manipulations, displays, and symbology can be found in Appendix A. Each of the four stages had a clearly defined primary task objective that could be performed manually or with the aid of automation. Each stage also had a secondary task of maintaining the flight parameters also shown in

Figure 19. The units listed are the commonly referenced units in flight environments and will be used throughout this thesis.



Stage	I	II	III	IV
Altitude (ft)	10,000	12,000	14,000	14,000
Airspeed (kts)	450	500	500	600
Distance to FEBA (nm)	37.5	15	5	Distance across FEBA 50

Figure 19. Graphical representation of a typical mission scenario.

Primary Tasks

Stage I

The primary task for Stage I was to identify eight high priority targets and add them to a shootlist, which was displayed on the HUD. The targets were, in order of priority; fire control radars, surface-to-air missile launchers, armored personnel carriers, and tanks. There were four groups of targets located in the engagement area (denoted by triangles in Figure 19). Two of the four target groups contained the highest priority targets. The other two groups may have contained some high level surface-to-air threats but the group, as a whole, did not contain the highest number of high priority threats. The engagement area was set up such that there were two near groups (closest to the ingress point) and two far groups (furthest from the ingress point). Pilots were instructed

to search the engagement area with a synthetic aperture radar (SAR) patch map and identify the best group of the two near groups that contained the highest number of high priority threats. Once this group had been identified, the pilot was instructed to place the highest priority threats from that one group on the shootlist. Pilots were informed that picking targets from both groups would negatively impact their ability to shoot the targets when they were in the engagement area. The pilots repeated this process for the two far groups of targets. In the manual condition, targets under the SAR patch map classified as a function of time under the SAR patch map. Automation in this stage facilitated the classification of the targets under the SAR patch map so that all targets were fully classified. Pilots were instructed to press a confirm button when the primary task was completed. If time ran out before the primary task was completed (reaching the next waypoint), attempts to add targets to the shootlist were prevented by the software.

Stage II

The primary task in Stage II was the prioritization of the targets on the shootlist that were selected in Stage I. The shootlist had eight slots available, one for each missile on the simulator ownship (see Appendix A). Ideally, at the end of Stage II the pilots chose four targets from the near group and ordered those in the first four slots from highest priority to lowest priority. The targets in the last four slots, from the far group, would be arranged in a similar manner. In manual mode, the pilots moved the targets up and down the shootlist using inputs from the stick and throttle, similar to a cut and paste function in a word processing program. In the automated condition, the targets were prioritized by the "in flight computer." Pilots were given the opportunity to change the order of the targets if they did not agree with the result of the automated sorting. Pilots accepted the order of the shootlist when they pressed the confirm button.

Stage III

The primary task in Stage III was to choose a flight path through the FEBA directed towards the targets picked in Stage I. There were four possible flight paths representing the 2×2 (near group $\times$ far group) combinations. In the manual condition the pilot toggled through the available flight paths (shown on the right HDD) until the path matched the ideal path for the targets on the shootlist. The automation in Stage III toggled to the best course. The pilot could change the course if needed and in either condition was asked to push the confirm button to accept the flight path displayed.

Stage IV

The primary task in Stage IV required the pilot to (a) avoid being shot down by surface-to-air missiles and (b) shoot the targets listed in the shootlist. In order to shoot a target, the pilot needed to be within range, within pitch and roll parameters, below a specified altitude, and designated on the target. In the manual condition, the pilot needed to monitor each of these parameters and assure they were within the specified tolerances before releasing the weapon. In the automated condition, a visual cue advised the pilot if they had achieved a proper firing solution and, if not, which parameter was not met. This cue was located on the HUD below the shootlist.

Design

Mission Scenarios

Initially, sixteen mission scenarios were created in ModSAF for data collection trials. Each ModSAF scenario corresponded to an experimental condition that resulted in the factorial combination of the two automation levels in each of the four mission stages (see Table 5). These sixteen scenarios represented low workload conditions. Adding four entities to each group in the sixteen scenarios created high workload conditions.

Thus, a high workload scenario was identical to its corresponding low workload scenario in every way except the number of entities in each group; six in the low workload condition and ten in the high workload condition.

Scenario	Stage I	Stage II	Stage III	Stage IV
1	Automated	Automated	Automated	Automated
2	Automated	Automated	Automated	Manual
3	Automated	Automated	Manual	Automated
4	Automated	Automated	Manual	Manual
5	Automated	Manual	Automated	Automated
6	Automated	Manual	Automated	Manual
7	Automated	Manual	Manual	Automated
8	Automated	Manual	Manual	Manual
9	Manual	Automated	Automated	Automated
10	Manual	Automated	Automated	Manual
11	Manual	Automated	Manual	Automated
12	Manual	Automated	Manual	Manual
13	Manual	Manual	Automated	Automated
14	Manual	Manual	Automated	Manual
15	Manual	Manual	Manual	Automated
16	Manual	Manual	Manual	Manual

Table 5. Experimental conditions.

Scoring

Primary Task Scoring (P)

A primary task score was computed for each of the mission stages and reflected the pilot's ability to perform the required primary tasks. Primary task points were awarded according to the following schedule:

Stage I – (320 points possible)

40 points for each high priority target in the highest priority group

20 points for a SA-15 or SA-9 not in the highest priority group

15 points for a ZSU not in highest priority group

10 points for any T-72 or BMP2

0 points if shootlist slot was left blank

Stage II – (160 points possible)

20 points for each target in correct position based on priority

All possible points were awarded for ties

Stage III – (150 points possible)

75 points for each correct course segment plotted

Stage IV – (320 points possible)

40 points for first missile launch on each target in the shootlist while meeting all launch conditions

20 points for additional missile launches on any target

0 points for missile launch while outside launch parameters

Secondary Task Scoring (S)

Secondary task scores were computed from deviations of commanded flight parameters. The component measures, respective units, and weighting factors are listed in Table 6. The weighting factors implemented were designed to equalize the impact of each measure due to the unit differences. Similar to primary task scores, secondary task scores were computed for each stage.

Altitude deviations were not calculated for a period of 15s while pilots were in a transition phase (ordered to go to a higher altitude). Similarly, airspeed deviations were not calculated for a period of 20s during transition phases. Heading deviations (Cross-track errors and Track angle errors) were calculated continuously as the heading did not change during stage transitions. Secondary task scores were not computed during Stage IV due to the unpredictability of pilot responses while engaging in evasive maneuvers in the event they were being shot at from surface-to-air missiles. Pilots were informed that the Stage IV score deviations would be suspended while they engaged in evasive maneuvers and would continue when the threat had been defeated.

Measure	Unit	Weighting Factor
Altitude	Feet	X 0.05
Airspeed	Knots	X 0.50
Cross-track	Nautical miles	X 5.00
Track error	Degrees	X 10.00

Table 6. Secondary Task weighting factors.

Total Scores (T)

Total scores were computed for each stage by subtracting secondary task scores from primary task scores (P – S = T). A Global Score (G) for each mission scenario was computed by subtracting the sum of all secondary task scores from the sum of all primary task scores as follows, $(P_I + P_{II} + P_{III} + P_{IV}) - (S_I + S_{II} + S_{III}) = G$.

Nomenclature

Table 7 lists the shorthand nomenclature that will be used for each of the variables listed. This shorthand will facilitate the rapid assessment of which variables are being referred to for the remainder of this document.

Variable			X	Y	Example
Independent Variables					
	Automation Level	X_Y	I / II / III / IV	A / M	I_A
	Workload Level	W_X	L / H		W_L
Dependent Variables					
	Primary Task Score	P_X	I / II / III / IV		P_I
	Secondary Task Score	S_X	I / II / III / IV		S_I
	Total Score	T_X	I / II / III / IV		T_I
	Global Score	G			

Table 7. Experimental nomenclature.

Experimental Design

A within-subjects design was employed in which a $2 \times 2 \times 2$ (current stage automation level × previous stage automation level × mission workload level) repeated-measures ANOVA was conducted on all stage scores, except the first stage, which had no previous stage. The Global Scores and subjective measures were analyzed by utilizing a 2×2 (Stage I automation level × mission workload level) repeated measures ANOVA. The analysis strategy employed for the Global Scores and subjective measures is consistent with and is most parsimonious for addressing the hypotheses stated in the introduction section for this experiment.

Procedure

Simulator Sickness Questionnaire

Pilots were asked to fill out a Simulator Sickness Questionnaire (Kennedy, Lane, Berbaum, & Lilienthal, 1993) prior to and immediately after any session that included time in the SIRE facility, regardless of the amount of time spent in the facility (see Appendix A). This ensured compliance with approved Institutional Review Board procedures that were put in place to identify potential ill effects associated with being in a virtual environment.

Training

Prior to the first training session, pilots were asked to familiarize themselves with the task requirements and cockpit switch operations outlined in the training manual (Appendix A) that was forwarded to them. Upon arrival, pilots were asked to fill out the informed consent sheets, a biographical information sheet, and were given a brief tour of the SIRE facility (see Appendix B). The training protocol utilized for this effort used both written training objectives and a training checklist (see Appendix C). The training

was divided into three functional areas; (a) simulator control, (b) switches, displays and tasks in manual mode without flying (static training), and (c) switches, displays, and tasks in automated and manual modes while flying (dynamic training). The simulator control objective that ensured each pilot familiarized themselves with the response characteristics of the simulator. This objective was usually met within a five-minute time frame.

The objectives in the second functional area of the training introduced the task objectives for each stage of the mission. The pilot was trained through verbal instruction from one experimenter while another experimenter checked off the items in the training checklist. The pilot was instructed on the use of the switches, display symbology and functionality, and the capabilities of these in relation to the stated primary and secondary task objectives. The pilot was also instructed to make errors so that corrective procedures could be demonstrated. The pilot was free to ask any questions during this and all phases of the training. Further, the pilot was able to go through this static training until they indicated they were ready to proceed to the next phase of the training.

The third phase of the training consisted of eight scenarios. The first two scenarios repeated the previous phase scenarios but, in addition, required the pilot to fly and maintain the flight parameters while performing the required tasks. In training scenarios three through six the automation was introduced to the pilot for each stage of the mission. Thus, scenario three contained an automated first stage while all other stages were performed manually. Likewise, scenario four contained an automated stage two while all other stages were performed manually, scenario five automated stage three and scenario six automated stage four. The training objective for these four scenarios was to point out the differences between the tasks when automation was present in the stage. The final two training scenarios were completed with all four stages in the automated condition to counterbalance the first two scenarios, in which all stages were completed in the manual condition.

The pilot was informed that they were free to request another static training trial before the data collection trials began. Data collection trials ensued only after the pilot indicated they were ready to proceed. The training phase of this experiment typically was conducted in four hours, in one or two separate training sessions. Data collection

trials were generally broken into separate sessions, over several days due to the length of the experiment. Pilots were required to participate in at least one static training trial before each new session began. The static training could be repeated until the pilot indicated they were ready to proceed with the data collection trials.

Data Collection

Upon the completion of training, each pilot participated in 32 data collection trials typically spread over two or three data collection sessions. At the end of each mission scenario, the pilot was instructed to fill out the on-screen subjective assessments of their mental workload, situation awareness, and their trust and confidence in the automation. These subjective measures are described in more detail below. Paper versions of these scales are located at the end of Appendix A. They input the values by moving the stick left or right and confirmed the value by moving a switch located on the throttle. Pilots were free to change any input up until the time they pressed the confirm button indicating that all input values were correct and they were ready to proceed with the next mission scenario.

Subjective Measures

Perceived Mental Workload

Subjective measures of workload were obtained utilizing the following six sub-scales of the NASA-Task Load Index (NASA-TLX): mental demand, physical demand, temporal demand, performance, effort, and frustration level (Hart & Staveland, 1988). Pilots were asked to rate (from 1-100) their perceived level of each of the sub-scales after each mission scenario.

Perceived Situation Awareness

The pilot's perceived situation awareness was measured after each mission scenario by utilizing the 3-D Situation Awareness Rating Technique or 3-D SART

(Taylor, 1990). The 3-D SART uses the three dimensions of attentional demand, attentional supply, and understanding. Pilots were instructed to rate each of the dimensions on a 1-7 scale. In addition to the three values associated with the 3-D SART, a question of overall situation awareness was also included in the survey.

Trust and Confidence

Again, after each mission scenario, pilots were asked to provide ratings on their trust in the automation using the Lee and Moray (1992) trust scale. In addition, a question asking the pilots to rate their confidence in their ability to complete the mission and a question asking the pilots to rate the reliability of the automation they encountered were included.

Experimental Debriefing

At the conclusion of all data collection trials the pilots were asked to fill out the debriefing questionnaire (Appendix D). Further, they were asked to provide detailed insights into the strategies they employed throughout the experiment.

Post-experimental trials

At the end of the debriefing session, each pilot was asked to help evaluate the training used in the experiment by participating in four additional trials. All pilots agreed to participate in the additional trials. After the trials were conducted, the pilot was informed that the data was going to be used to compare the four just completed trials to the trials in the data collection session that matched the experimental conditions.

Results

Analysis Strategies

For reasons of continuity, the Stage dependent results are listed and grouped by the Stages. Thus, all of the Stage I dependent variables are discussed first followed by Stage II and so forth. A summary of the results appears at the beginning of each section. The presentation of results within each Stage is in accordance with the methodology for computing the scores. The Primary Task Score (P) is listed first followed by the Secondary Task Score (S) then the Total Score (T) for each Stage. After the Stage dependent variables are discussed the mission dependent measures are presented. The mission dependent variables were computed and/or derived after the mission had been completed and are representative of the whole mission and not any particular Stage. The Global Score (G) for performance is presented first, followed by the results of the subjective ratings, training assessment, and a synopsis of the de-briefing information. For reference, all ANOVA tables for the performance and subjective measures are listed in Appendix E.

There were a number of significance tests carried out for all of the dependent measures. A general concern regarding this approach is the control of Type I errors, concluding there is a relationship between two variables when in fact there is no such relationship present. Two possible causes for Type I errors in a repeated measures design are violations of homogeneity and violations of sphericity. To guard against these violations, the Huynh-Feldt adjustment was utilized in all omnibus comparisons conducted. Further, an alpha level of .05 was adopted for all omnibus tests.

Another statistical concern with conducting multiple comparisons is the control of familywise and experimentwise error rates. Although pre-planned comparisons can be utilized to reduce the likelihood of these errors, a more conservative approach is often more appropriate when a multitude of comparisons are possible. The temporal nature of this experiment (Stage I before Stage II before Stage III etc.) limits the utility of

conducting all possible comparisons. It was decided a priori that the Bonferonni adjustment to the alpha level for post-hoc comparisons would apply only to the potential comparisons of interest. Thus the post-hoc comparison alpha levels used were determined by dividing .05 by the logical number of comparisons within these constraints. The alpha level for individual cell mean comparisons was determined by dividing .05 by the total number of comparisons possible.

Performance Measures

Stage I Measures

Summary of Stage I Measures

During the first stage of the mission, pilots who were aided by automation were able to perform the required primary task of identifying high priority targets and placing them on the shootlist faster and with more accuracy than when they performed those tasks manually. Further, the automation in Stage I enabled the pilots to more closely adhere to the commanded flight parameters. In general, the workload level exacerbated these effects in the manual condition with high workload conditions engendering the lower performance ratings.

Stage I Primary Task Scores

All Stage I dependent measures were analyzed using a 2 × 2 repeated measures ANOVA. The first factor was the level of automation in Stage I (manual, automated) and the second factor was the workload level (low, high). For the Primary Task in Stage I (P_I), identifying and adding targets to the shootlist, there was a significant main effect for Stage I automation level, $\underline{F}(1,7) = 104.14$, $\underline{p} < .05$, and workload level, $\underline{F}(1,7) = 19.26$, $\underline{p} < .05$, as well as the two-way interaction between the Stage I automation level and the workload level, $\underline{F}(1,7) = 17.95$, $\underline{p} < .05$. As seen in the Figure 20, the P_I under the

automated Stage I condition did not differ between high and low levels of workload while the manual Stage I condition was influenced by workload level.

Subsequent post-hoc comparisons of the two-way interaction revealed that the largest disparity (72.19 points) was the significant difference, $\underline{F}(1,7) = 91.42$, $\underline{p} < .0125$, in the high workload condition between Stage I automated ($\underline{M} = 307.89$, $\underline{SE} = 4.26$) and manual levels ($\underline{M} = 235.70$, $\underline{SE} = 7.35$). The second largest difference (45.24 points) was due to the significant difference, $\underline{F}(1,7) = 35.90$, $\underline{p} < .0125$, in the manual condition between high ($\underline{M} = 235.70$, $\underline{SE} = 7.35$) and low ($\underline{M} = 280.94$, $\underline{SE} = 6.41$) workload levels. The low workload condition also provided a significant difference, $\underline{F}(1,7) = 12.74$, $\underline{p} < .0125$, between Stage I automated and manual levels.

Overall, this suggests that pilot performance on the primary task increased when Stage I was automated regardless of the workload level. Further, when Stage I was performed manually, not only were the primary scores lower as a group, they were differentially lower with the lowest score being engendered by the high workload level. Moreover, it should be noted that there was virtually no difference between the scores in the Stage I automated condition across workload levels, as illustrated in Figure 20. This suggests that the automation in Stage I nullified the effects of workload as compared to when the pilots had to perform Stage I manually.

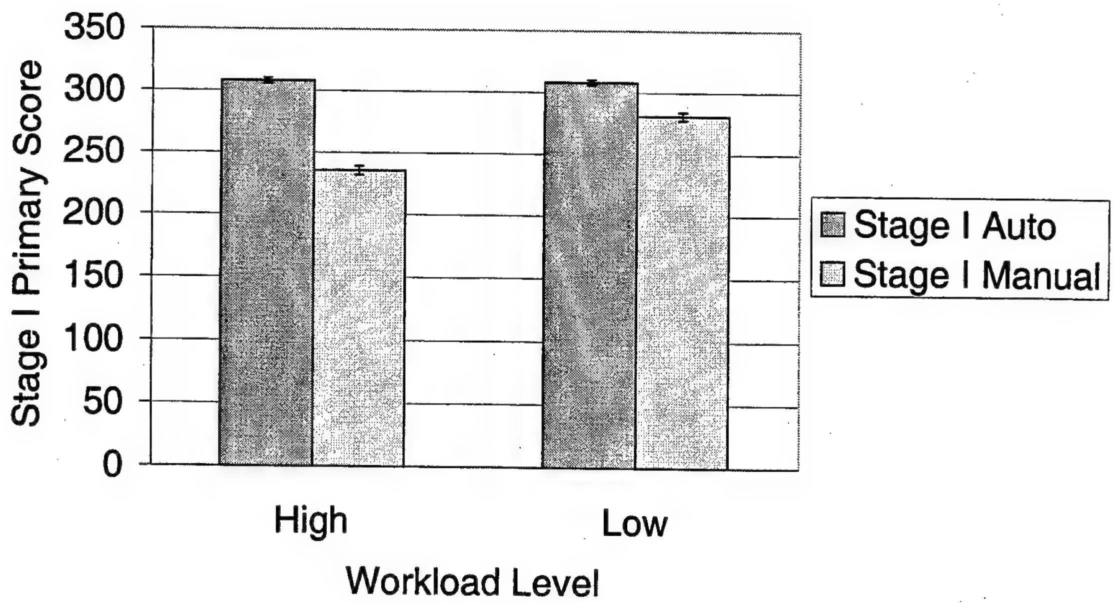


Figure 20. Stage I primary task scores (maximum 320) as a function of workload level and Stage I automation level. Error bars represent one standard error of the mean.

Stage I Secondary Task Scores

The Stage I Secondary Task Scores (S_I), a composite of altitude and airspeed deviations and track and cross-track deviations in Stage I, were submitted to the same 2 × 2 (Stage I automation level × workload level) analysis as previously reported. Stage I automation level and workload main effects were not significantly different for the omnibus test ($p > .05$). The two-way interaction between these factors was significant, $F(1,7) = 12.11$, $p < .05$, and is illustrated in Figure 21. The post-hoc comparisons for the interaction indicate that one source of variance is the significant difference between the automated and manual conditions in Stage I at the low workload level, $F(1,7) = 41.56$, $p < .0125$. The average difference of 10.56 points between the automated-low ($M = 48.65$, $SE = 4.56$) and the manual-low ($M = 38.09$, $SE = 3.40$) conditions is not practically large but it is in the opposite direction than expected. The pilots maintained their flight parameters with less integrity

when the automation was present and workload was low. This unexpected finding will be addressed below in the section that addresses the de-briefing data. The other significant source of variance for this interaction was the difference within the automated condition between the high ($\underline{M}$ = 38.31, $\underline{SE}$ = 3.29) and low ($\underline{M}$ = 48.65, $\underline{SE}$ = 4.56) levels of workload, $\underline{F}(1,7) = 39.87$, $\underline{p} < .0125$. The other two possible sources of variance failed to differ in any substantial manner ($\underline{p} > .05$).

The Stage I Secondary Task Score (S_I) is a composite of deviations from the commanded altitude, airspeed, track angle, and cross-track errors. These four, as a group, represent the secondary task imposed on the pilot during the mission. A composite score is more than sufficient in determining potential trade-offs between primary and secondary task performance and is thus reported for each stage. However, the individual factors may be operationally important and therefore ANOVA tables of these individual flight parameters are listed in Appendix E for reference.

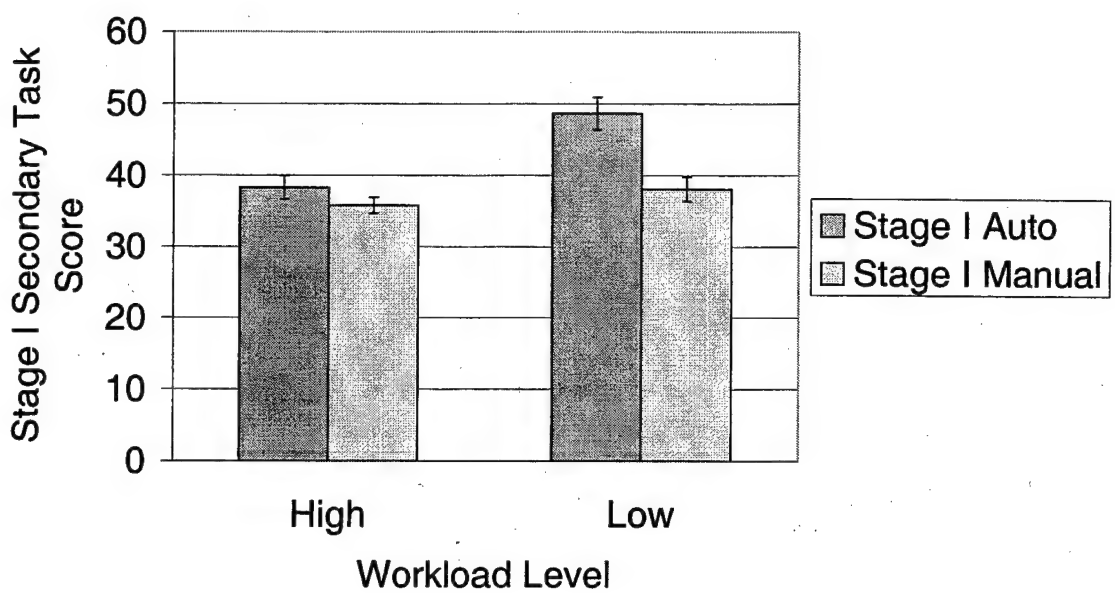


Figure 21. Stage I Secondary Task scores as a function of workload level and Stage I automation level. Error bars represent one standard error of the mean.

Stage I Total Scores

A similar 2 × 2 repeated measures ANOVA was conducted on the Total Stage I scores (T_I), which reflected the net difference of $P_I - S_I$. The omnibus test revealed that there was a significant difference for the main effect of Stage I automation level, $F(1,7) = 31.19$, $p < .05$, and the two-way interaction between Stage I automation level and workload level, $F(1,7) = 20.65$, $p < .05$. The two-way interaction is illustrated in Figure 22. Subsequent post-hoc comparisons revealed that the difference between the I_A-W_H ($M = 269.58$, $SE = 6.35$) and the I_M-W_H ($M = 199.89$, $SE = 7.65$) conditions was the largest source of significant variance between the marginal means, $F(1,7) = 70.63$, $p < .0125$. There was also a significant difference between the high ($M = 199.89$, $SE = 7.65$) and low ($M = 242.85$, $SE = 8.00$) workload conditions when Stage I was completed manually, $F(1,7) = 26.83$, $p < .0125$.

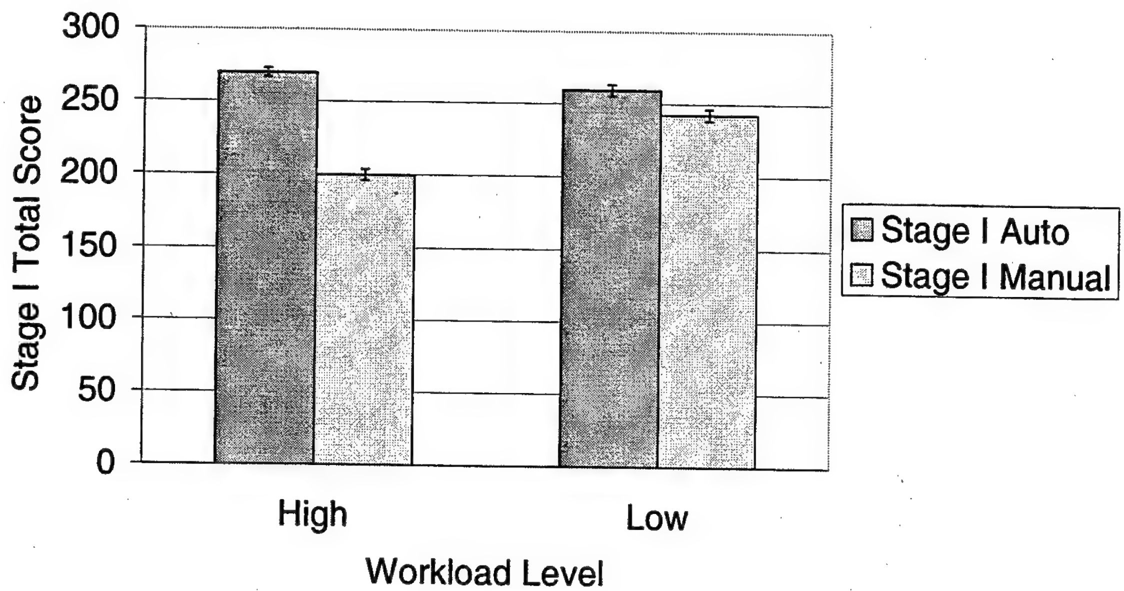


Figure 22. Stage I Total scores as a function of workload level and Stage I automation level. Error bars represent one standard error of the mean.

The Total Stage I scores summarize pilot performance on the primary and secondary tasks. On inspection, the Total Stage I score interaction between Stage I automation level

and workload looks similar to the Primary Task score interaction for Stage I. The difference is that the Stage I Secondary Task score has attenuated the Stage I Primary Task score and thus the Stage I Total score reflects the synergistic relationship between primary and secondary task performance. From this perspective, a stronger case may be made about which conditions engendered better overall performance with respect to the mission goals. The mission objectives were better met under the automated Stage I condition. Pilots were able to identify and add more high priority targets to the shootlist and maintain the flight parameters more closely when aided by the automation. There was a slight impairment (10.34 points) to the Total Stage I score in the automated condition between the high (M = 269.58, SE = 6.36) and low (M = 259.24, SE = 7.87) workload levels indicating that the secondary task load influenced the I_A-W_L condition more than the I_A-W_H condition. This comparison is based on the lack of a significant difference in the automated Stage I condition between high and low workload levels in the Primary Task Stage I score ($p > .05$). It is equally important to note the effects of the workload manipulation on the mission objectives. In the low workload condition pilots were reasonably able to effectively search and add targets to the shootlist as well as maintain the commanded flight parameters in both the automated and manual conditions. However, when pilots had more targets in the environment (high workload condition), their ability to search the area, identify appropriate targets, and add them to the shootlist while maintaining the flight parameters was hindered in the Stage I manual condition as compared to the automated condition.

Stage I Other Measures

Pilots pushed the confirm button to record the time since the beginning of the stage when they thought they had completed all of the tasks assigned to them in that stage. This measure is decidedly different than the performance measures in that it reflects the temporal aspect of the task and not the accuracy of the task. There was a significant difference for the main effect of Stage I automation level, $F(1,7) = 32.55$, $p < .05$, as well as the two-way interaction between Stage I automation level and workload, $F(1,7) = 21.54$, $p < .05$. This interaction, shown in Figure 23, illustrates the results of the post-hoc comparisons performed

for the analysis. Of the four possible comparisons, three were significant; the difference in the high workload condition across both Stage I automation levels, $\underline{F}(1,7) = 305.00$, $\underline{p} < .0125$, the difference in the low workload condition across both Stage I automation levels, $\underline{F}(1,7) = 118.81$, $\underline{p} < .0125$, and the difference in the manual condition across both workload conditions, $\underline{F}(1,7) = 23.13$, $\underline{p} < .0125$. Combined, these results reveal that pilots were confident they were through completing the necessary primary tasks at a minimum of 22.11s earlier when Stage I was automated.

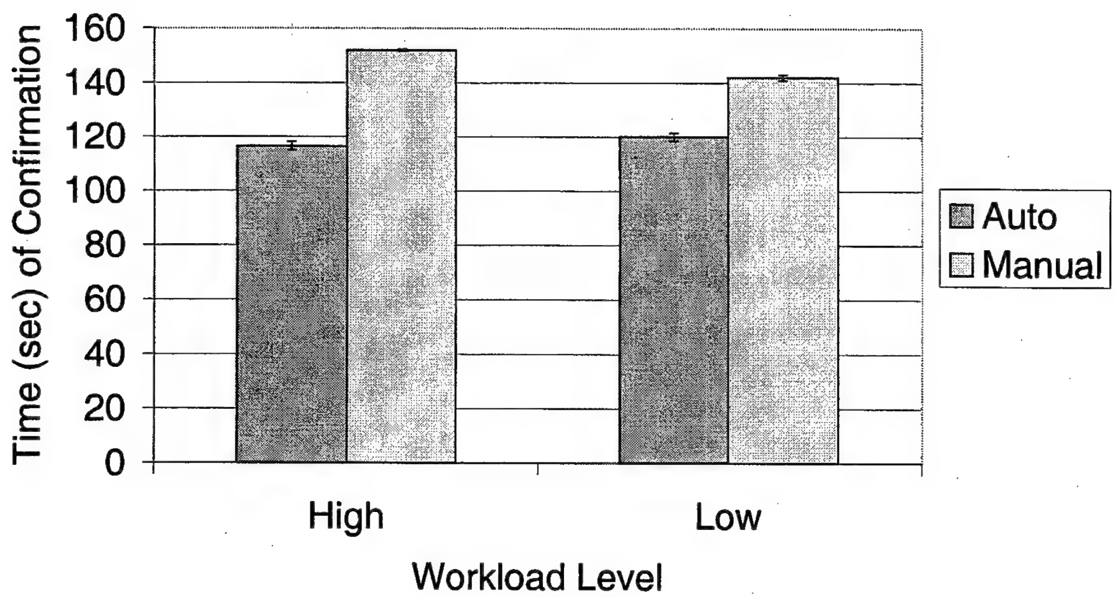


Figure 23. Average time (s) the pilots pressed the confirmation button in Stage I as a function of workload level and Stage I automation level. Error bars represent one standard error of the mean.

A potential criticism of using the confirmation button press as an indication of temporal variance is the inherent cognitive process time included by the pilot to evaluate the state of task completion. To address this concern, the time of the last switch action made by the pilot was also recorded. This measure is sensitive to the function of performing the task without the cognitive evaluation of its completeness. This dependent measure was subjected to the same 2 × 2 repeated measures ANOVA for Stage I. The omnibus test revealed that there was a significant main effect for Stage I automation level, $\underline{F}(1,7) = 63.55$, $\underline{p} < .05$, and workload level, $\underline{F}(1,7) = 6.52$, $\underline{p} < .05$. The two-way interaction of these factors also produced a significant difference, $\underline{F}(1,7) = 24.34$, $\underline{p} < .05$, for the time of the last switch action in Stage I. The pattern of the interaction was identical to that of the confirm time analysis. The same three post-hoc comparisons were similarly significantly different (all $\underline{p} < .05$) indicating that the last switch action coincided with the confirm times for these factors.

The last measure to consider in the Stage I analysis is the lapsed time spent in the stage. As stated previously, the stage transition points were waypoint dependent. Given that the pilot had to fly at the commanded flight parameters, it should be true that they should all spend the same amount of time in each stage. There is a high level of confidence that one of the subtractor measures would account for any variance if the pilot deviated from one of the parameters in order to have more time available to complete the primary tasks required. For example, by decreasing airspeed or oscillating around the flight path, the pilot could increase the time it takes to get to the next waypoint. To test this potential strategy, an analysis of the time spent in the stage was conducted. The notion is that there should not be any difference in the time spent in any stage regardless of the experimental conditions imposed. The 2×2 ANOVA revealed that the amount of time spent in Stage I was not significantly different for any of the experimental conditions ($p > .05$) indicating that the pilots did not try to hedge the flight parameters to their benefit.

Stage II Measures

The analyses of Stage II dependent measures were similar to that of the Stage I measures but included the automation level in Stage I as an additional factor. This approach was similarly adopted for all remaining Stage factors. Thus, analyses for Stage II, Stage III, and Stage IV were conducted with a $2 \times 2 \times 2$ (previous stage automation level $\times$ current stage automation level $\times$ workload level) repeated-measures ANOVA. The primary reason for adopting this approach was the continuous-time framework of the mission. Each stage, except Stage I, was preceded by another stage, which may or may not have influenced the performance in the current stage. This approach, based on Markov chain decision-making, allows the examination of the current state and the previous state but limits the effects of influence to these two stages.

Summary of Stage II Measures

In general, the Stage II dependent measures were primarily affected by the automation level in the previous Stage I. When Stage I was automated, Stage II performance was elevated as indicated by the overall metric which takes into account the primary and secondary task performance combined.

Stage II Primary Task Scores

The primary task in Stage II was the sorting of the targets on the shootlist that were identified and added to the shootlist during Stage I. Targets were sorted based on their location (near-group or far-group) and threat potential rank (SA-6 FCR, SA-6 TEL, SA-15 etc.). The omnibus test revealed that there was a significant difference in the Primary Task scores for the main effect of Stage I automation level, $F(1,7) = 7.72$, $p < .05$, and workload level, $F(1,7) = 29.16$, $p < .05$. Further, the analysis revealed that there was a significant two-way interaction between the levels of workload and Stage I automation levels for the Stage II Primary Task scores, $F(1,7) = 7.00$, $p < .05$. Post-hoc comparisons for this interaction, shown in Figure 24, include the significant 14.07 point difference, $F(1,7) = 36.17$, $p < .0125$, in the high workload condition when the previous Stage I was automated ($M = 154.69$, $SE = 2.41$) compared to when it was performed manually ($M = 140.62$, $SE = 4.03$). Further, the comparisons revealed that the 11.56 point difference between high and low workload levels was significantly different when the pilots performed the previous Stage I manually, $F(1,7) = 24.45$, $p < .0125$.

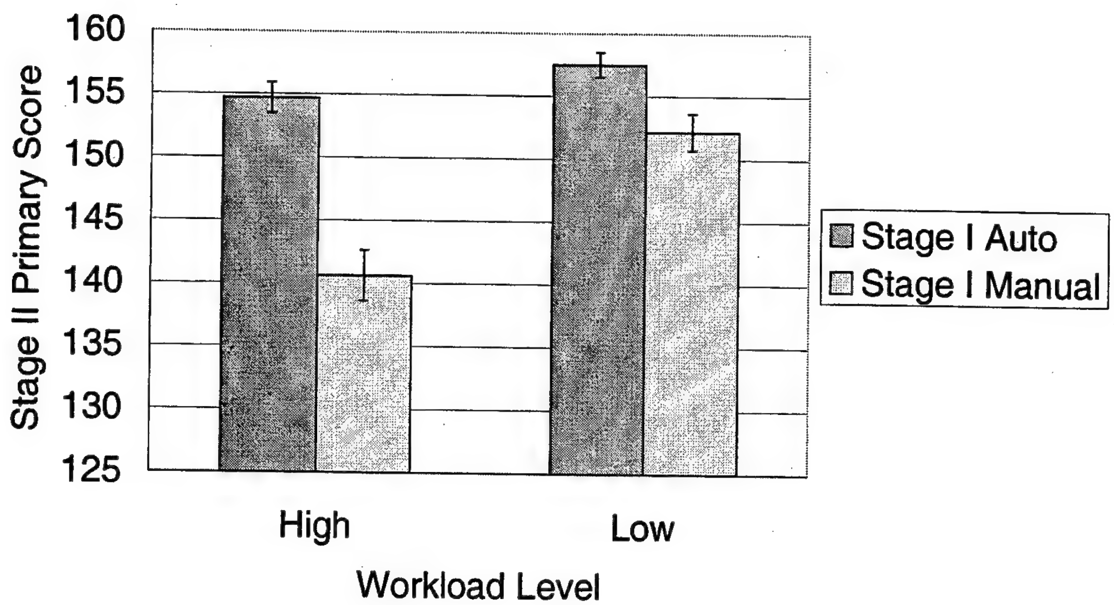


Figure 24. Primary scores in Stage II as a function of workload level and Stage I automation level. Error bars represent one standard error of the mean.

The analysis did not reveal a significant difference for the main effect of Stage II automation level or any other interactions for the Primary Task Score in Stage II ($p < .05$). This indicates that the majority of variance seen in the Stage II Primary Task scoring was due to Stage I automation level and the workload level. Pilots were better able to sort the shootlist after completing an automated Stage I. Further, their general performance of the Stage II task was hindered when the pilots transitioned from a manual Stage I condition. A high workload level further diminished pilots' ability to complete this task effectively.

Stage II Secondary Task Scores

The Secondary Task scores for Stage II were not significantly different for any of the three factors tested ($p > .05$) indicating that the pilots were generally able to stay within the prescribed flight parameters during the second stage regardless of the experimental conditions. There were, however, significant differences in the individual component

measures of altitude and airspeed deviations (significant main effect for Stage I automation level), and track angle deviations (significant main effect for Stage II automation level). These differences are outlined in Appendix E.

Stage II Total Scores

The Total Stage II scores were not significantly affected by the level of automation in Stage II or the workload level of the mission according to the omnibus test conducted ($p >$ .05). The Total Stage II scores were, however, significantly different depending on the main effect of Stage I automation level, $F(1,7) = 7.94$, $p < .05$, and are depicted in Figure 25. Total Stage II points were 21.82 higher when preceded by an automated Stage I (M= 110.35, SE = 4.23) than when preceded by a Stage I that was performed manually (M = 88.53, SE = 4.15). There were no other significant sources of variation for this dependent measure ($p >$ .0125).

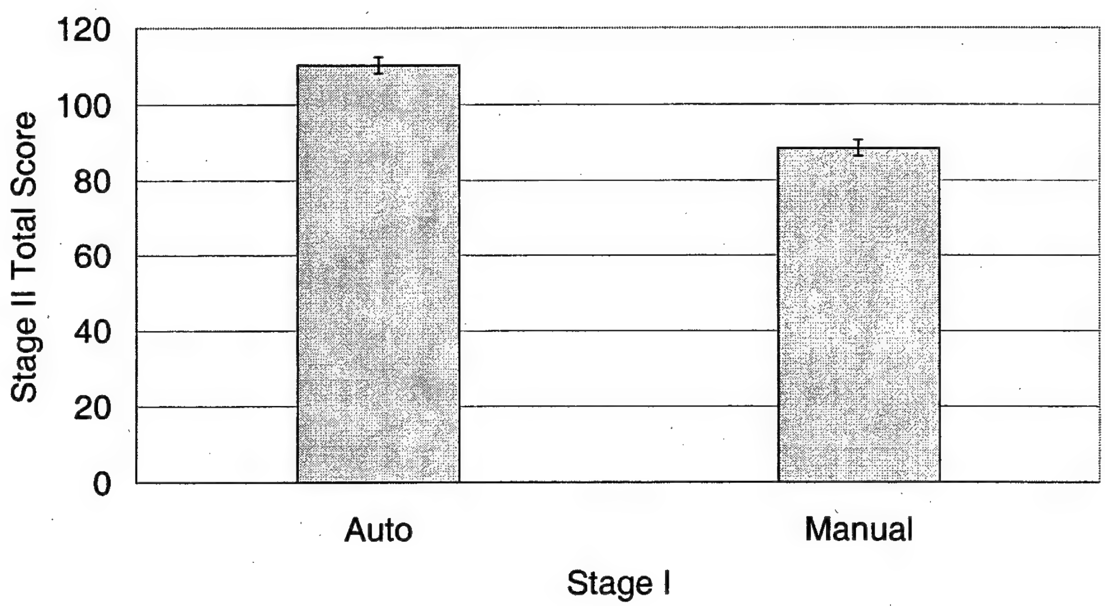


Figure 25. Total Stage II scores (maximum 120) as a function of Stage I automation level. Error bars represent one standard error of the mean.

Stage II Other Measures

The pilots' assessments of when they completed all the required tasks in Stage II, indicated by the time the confirm button was pressed, was significantly different for the main effect of workload level, $\underline{F}(1,7) = 10.37$, $\underline{p} < .05$. On average, pilots pressed the confirm button 4.8 s faster when the mission workload level was low ($\underline{M} = 32.73$, $\underline{SE} = 1.16$) compared to when the mission workload was high ($\underline{M} = 37.53$, $\underline{SE} = 1.21$). The last switch action in Stage II mirrored this significant difference, $\underline{F}(1,7) = 5.76$, $\underline{p} < .05$, showing an average 4.26s advantage when the mission workload level was low ($\underline{M} = 21.02$, $\underline{SE} = 1.26$) compared to when it was high ($\underline{M} = 25.28$, $\underline{SE} = 1.38$). Similar to the Stage I result, there was not an appreciable difference in the total amount of time the pilots spent in Stage II among all of the experimental conditions ($\underline{p} > .05$).

Stage III Measures

Summary of Stage III Measures

In summary, the Stage III performance metrics were affected by the interaction of the previous (Stage II) automation level, the current (Stage III) automation level, and the level of workload experienced.

Stage III Primary Task Scores

The primary task for Stage III was to choose the flight path through the FEBA that matched the groups of targets that were on the shootlist. Of the 150 points possible, 75 were awarded if the pilots chose the correct near-group path and 75 points for the far-group path. Although there was not a significant main effect for any of the three main factors evaluated in the omnibus test, there was a significant two-way interaction between the Stage II automation level and the Stage III automation level, $\underline{F}(1,7) = 9.33$, $\underline{p} < .05$. Post-hoc marginal mean comparisons revealed a significant difference between the Stage II automated and manual conditions when Stage III was manual, $\underline{F}(1,7) = 10.50$, $\underline{p} < .05$. On average, pilots scored 7.03 points higher in the Stage III manual condition after transitioning from an

automated Stage II condition (M = 150.00, SE = 0.00) than from a manual Stage II condition (M = 142.97, SE = 2.75). It should be noted that the II_A-III_M condition, collapsed across workload level, engendered a perfect score by all pilots. Because of this, the standard error of the mean for this condition is zero. The previous comparison was conducted utilizing a pooled error term that resulted in a pooled standard error of the mean (SE = 1.53) for all of the conditions. The other three post-hoc comparisons conducted for the Primary Task scores for Stage III did not yield a difference that was statistically significant ($p > .0125$). Likewise, neither of the other two-way interactions nor the three-way interaction reached significance ($p > .05$) for the Primary Task scores for Stage III.

Stage III Secondary Task Scores

The omnibus ANOVA of the Secondary Task Stage III scores produced significant differences for the two-way interaction between Stage II automation levels and Workload levels, $F(1,7) = 5.87$, $p < .05$, and the three-way interaction between the automation levels in Stages II and III and the mission workload levels, $F(1,7) = 6.78$, $p < .05$. There are three possible ways to interpret a three-way interaction. Due to the temporal nature of the factors in this experiment, the most logical explanation of this three-way interaction is that the two-way interaction between the mission workload level and the Stage III automation level was different for the two proceeding Stage II automation levels.

The three-way interaction, shown in Figure 26, demonstrates that when pilots transitioned from an automated Stage II to an automated Stage III they were less able to maintain the flight parameters under the low workload condition compared to the high workload condition as indicated by the higher Stage III Secondary Task scores. When pilots transitioned from a manual Stage II condition to an automated Stage III they had more trouble maintaining their flight parameters under the high workload condition compared to the low workload condition. While the simple interaction effects did not attain a significant difference ($p > .025$), the post-hoc comparisons of the cell means confirmed this conclusion. There was a significant difference between the II_A-III_A-W_H and II_A-III_A-W_L conditions, $t(7) = 3.43$, $p < .0125$, as well as the II_M-III_A-W_H and II_M-III_A-W_L conditions, $t(7) = 3.46$, $p < .0125$.

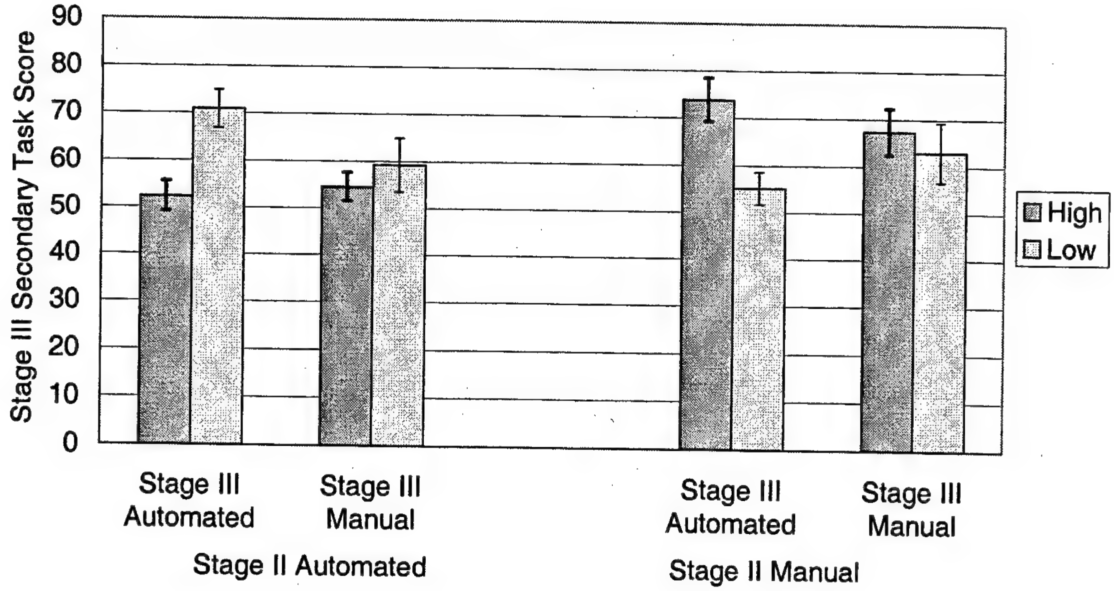


Figure 26. Secondary Task Stage III scores as a function of workload level, Stage II automation level, and Stage III automation level. Error bars represent one standard error of the mean.

Stage III Total Scores

The omnibus ANOVA of the Total Stage III scores also produced a significant two-way interaction for Stage II automation levels and Workload levels, $\underline{F}(1,7) = 5.78$, $\underline{p} < .05$, as well as the three-way interaction between workload, Stage II, and Stage III automation levels, $\underline{F}(1,7) = 36.16$, $\underline{p} < .05$. The three-way interaction for the Total Stage III scores is depicted in Figure 27. Note that a direct comparison between this and the previous figures is not appropriate because the Secondary Task score scale is opposite the Total score scale. In the previous figure, higher scores indicated less ideal performance. In the present figure, a higher score indicates a higher level of proficiency.

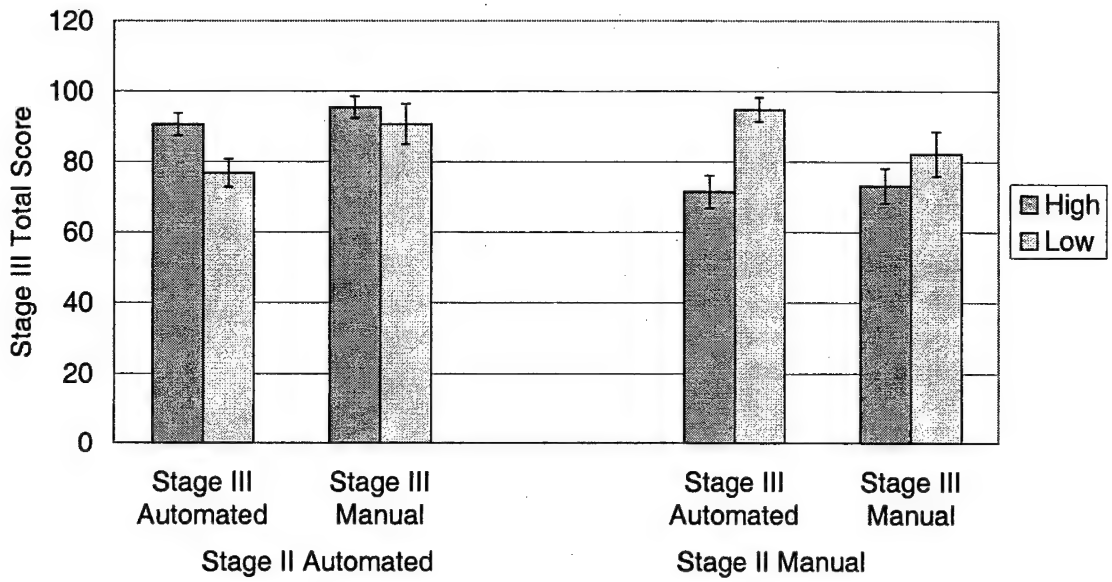


Figure 27. Total Stage III scores shown as a function of workload, Stage II automation level and Stage III automation level. Error bars represent one standard error of the mean.

The simple interaction effect for automated Stage II conditions was significantly different, $\underline{F}(3,7) = 34.48$, $\underline{p} < .025$, indicating that the four conditions where pilots transitioned from an automated Stage II were substantially different from each other. The subsequent post-hoc cell mean comparisons revealed significant differences between the II_A-III_A-W_H condition and the II_A-III_A-W_L condition, $\underline{t}(7) = 7.11$, $\underline{p} < .0125$, as well as the II_A-III_A-W_L condition and the II_A-III_M-W_L condition, $\underline{t}(7) = 7.17$, $\underline{p} < .0125$. The simple interaction test for the four conditions in which pilots transitioned from the Stage II manual condition was also significantly different, $\underline{F}(3,7) = 60.39$, $\underline{p} < .025$. Comparisons of the cell means revealed a significant difference between the II_M-III_A-W_H and II_M-III_A-W_L conditions, $\underline{t}(7) = 12.01$, $\underline{p} < .0125$, the II_M-III_M-W_H and II_M-III_M-W_L conditions, $\underline{t}(7) = 4.65$, $\underline{p} < .0125$, and the II_M-III_A-W_L and II_M-III_M-W_L conditions, $\underline{t}(7) = 6.45$, $\underline{p} < .0125$.

The findings for this three-way interaction can be best explained, again, by starting at the transition from Stage II. If pilots transitioned from an automated Stage II they completed the primary and secondary tasks, on average, better when the Stage III was manual under both levels of workload. If Stage III was automated however, workload did become a significant factor in determining how well they performed these tasks, with low levels of workload engendering a lower Total Stage III score. If the pilots transitioned from a Stage II manual condition the previous pattern of performance was reversed. The pilots did not perform the primary and secondary tasks better when Stage III was performed manually. Instead, they performed better when Stage III was automated, but only in the low workload condition. The same transition in the high workload condition, on average, produced the lowest Total Stage III scores. Further, the low workload condition showed an overall Stage III decrement when transitioning from a Stage II manual condition between Stage III automated and manual conditions whereas there was a performance benefit seen between the Stage III automated and manual conditions after transitioning from a Stage II automated condition. That same pattern was not evident under high workload conditions. Overall mission objectives in Stage III were better met when the pilots transitioned from an automated Stage II compared to a manual Stage II.

Of interest is a lack of a significant two-way interaction between Stage II and Stage III automation levels ($p > .05$) in the Total score when the same interaction was found for the Primary Task score in Stage III. This indicates that the primary task score differences were not pervasive enough to prevent mediation from the Secondary Task scores.

Stage III Other Measures

The pilot perception of when they completed the required tasks in Stage III was not significantly different between any of the conditions. The last switch action in Stage III was significantly different for the main effect of workload level, $F(1,7) = 12.63$, $p < .05$, and the two-way interaction between Stage II and Stage III automation levels, $F(1,7) = 12.79$, $p < .05$. This interaction, shown in Figure 28, does not appear to be of practical importance until the lack of any significant differences ($p > .05$) in the confirm button presses are taken into

consideration. This suggests that the differences in this interaction represent the additional time the pilots needed to assure that they had completed the required primary tasks in Stage III.

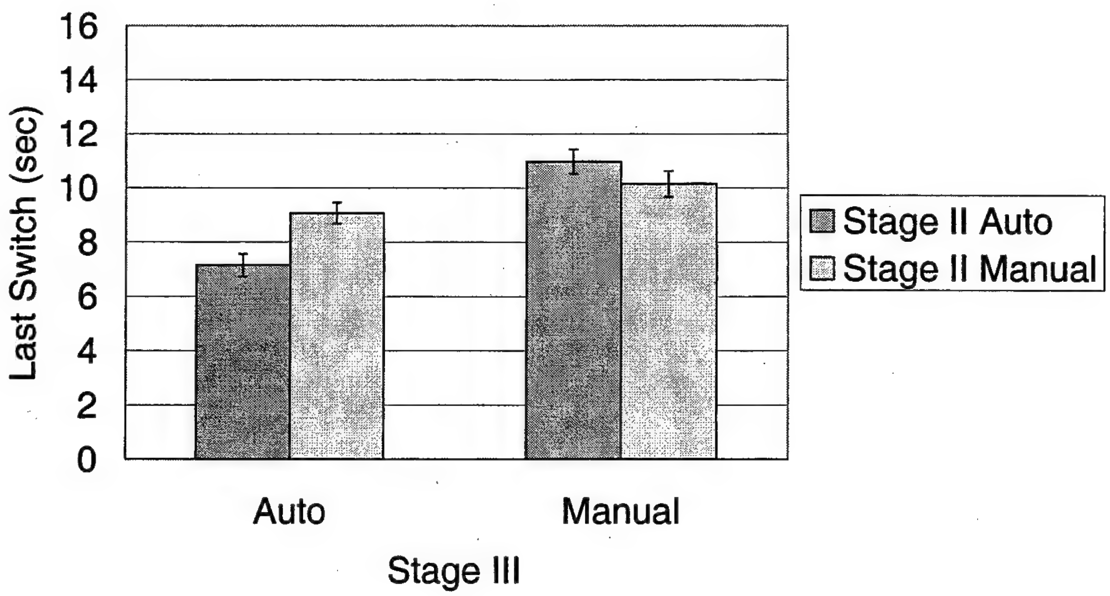


Figure 28. Time of the last switch action (s) in Stage III as a function of Stage II automation levels and Stage III automation levels. Error bars represent one standard error of the mean.

The comparison for this interaction revealed that there was a significant time delay in the last switch action after transitioning from an automated Stage II condition between the automated Stage III condition (M = 7.15, SE = 0.84) and the manual Stage III condition (M = 10.97, SE = 0.89), $F(1,7) = 49.68$, $p < .0125$. Further, there was a significant difference in the Stage III automated condition, $F(1,7) = 12.49$, $p < .0125$, depending on if the transition from Stage II was from a manual condition (M = 9.07, SE = 0.78) or an automated condition (M = 7.15, SE = 0.84).

As with the last two stages, there was not an appreciable difference in the amount of time the pilots spent in Stage III across the experimental conditions ($p > .05$). Again, this indicates that the pilots did not implore a strategy of maximizing the allowable deviations in the flight parameters to gain an advantage in completing their primary tasks.

Stage IV Measures

Summary of Stage IV Measures

Similar to the previous summaries, Stage IV performance was affected by the previous interaction of Stage III automation level and the current Stage IV automation level. Automation in both stages lead to better performance in the primary task of shooting the targets in the engagement area. Further, the automation allowed the pilots to shoot the targets faster on average.

Stage IV Primary Task Scores

The primary task in Stage IV was to shoot the targets that were identified and added to the shootlist in Stage I and sorted according to priority in Stage II. The secondary task in Stage IV was to follow the flight plan selected in Stage III at the pre-assigned altitude and airspeed. The primary scores for Stage IV were analyzed by including the Stage III automation level, the Stage IV automation level, and the level of workload for the mission. The main effects did not reach a level of statistical significance for the omnibus test performed ($p > .05$). There was a significant two-way interaction between Stage III and Stage IV automation levels for the Primary Task scores in Stage IV, $F(1,7) = 7.45$, $p < .05$. Comparisons conducted for this interaction (see Figure 29) revealed that the source of variance derived from the significant difference in the Stage III automated condition between Stage IV automation levels, $F(1,7) = 11.30$, $p < .0125$. The other comparisons conducted did not reach a statistically significant difference ($p > .0125$). As illustrated in Figure 29, Primary Task Stage IV scores did not differ significantly when pilots transitioned from a Stage III manual level, regardless of the automation level experienced in Stage IV. However

when pilots transitioned from an automated Stage III the Primary Task Stage IV score differed depending on the automation level experienced in Stage IV.

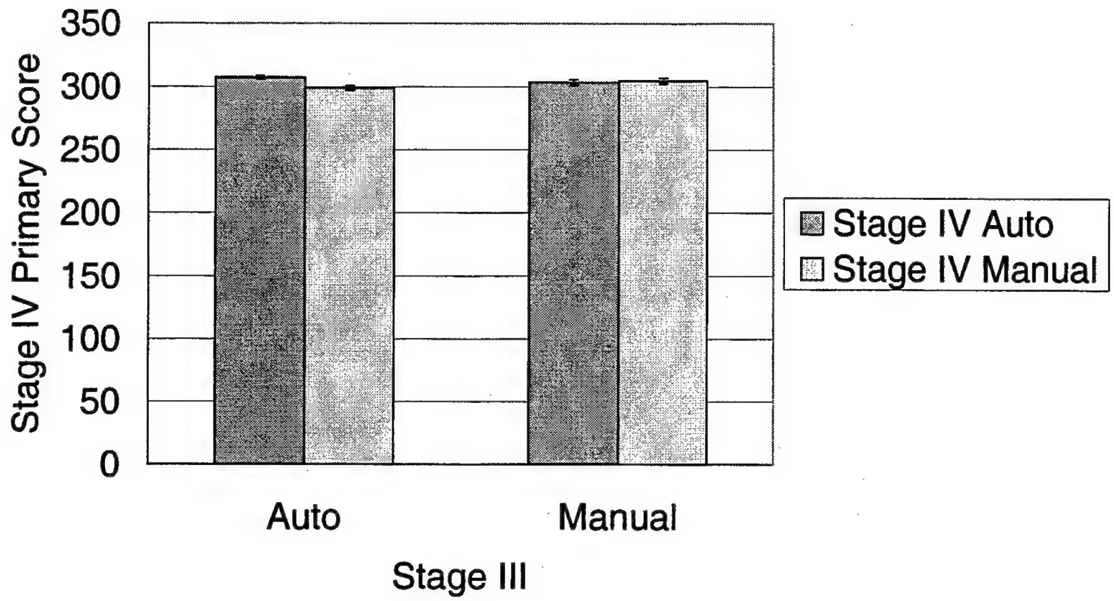


Figure 29. Primary score for Stage IV as a function of Stage III automation levels and Stage IV automation levels. Error bars represent one standard error of the mean.

Stage IV Secondary Task Scores

As previously stated there were no secondary task scores assigned to Stage IV. The pilots were instructed to follow the route they selected in Stage III and maintain their assigned altitude and airspeed. Additionally, they were instructed that the maintenance of these parameters was not as important as avoiding any surface-to-air armaments or shooting the targets that were listed on their shootlist.

Although Stage IV does not include secondary task measures, it does provide other measures that are unique to pilot performance in Stage IV. These measures reflect the amount of time the pilot took to fire the missiles at the designated targets on the shootlist. There are three related measures that capture this performance: time to shoot the near group

of targets, time to shoot the far group of targets, and the time to shoot at the two groups averaged across groups. The measures are referred to as Group 1 Range, Group 2 Range, and Average-Range, respectively. The range was measured from the time the first shot was taken to the time the last shot was taken for each target group. In the following analyses, the scenarios that did not have eight targets on the shootlist and those that deviated from having four targets from each of the two groups were omitted. In all, 28 of the 256 trials met that criteria and were excluded.

Stage IV Average-Range

The omnibus ANOVA revealed that there was a significant difference in the time to shoot the missiles, averaged between the near and far target groups, for the main effect of Stage III automation level, $\underline{F}(1,7) = 19.09$, $\underline{p} < .05$, Stage IV automation level, $\underline{F}(1,7) = 19.06$, $\underline{p} < .05$, and the two-way interaction between Stage III and Stage IV automation levels, $\underline{F}(1,7) = 8.34$, $\underline{p} < .05$. The two-way interaction for the average range measure is depicted in Figure 30.

On average, pilots were able to get all of the required shots fired 2.78s faster when they transitioned from an automated Stage III condition ($\underline{M} = 14.85$, $\underline{SE} = 0.56$) over a manual Stage III condition ($\underline{M} = 17.63$, $\underline{SE} = 0.58$). Tests of the simple interaction revealed that there was a significant difference in the range of time pilots took to shoot all targets, $\underline{F}(1,7) = 51.24$, $\underline{p} < .025$, after transitioning from an automated Stage III condition between an automated Stage IV condition ($\underline{M} = 12.16$, $\underline{SE} = 0.56$) and a manual Stage IV condition ($\underline{M} = 17.53$, $\underline{SE} = 0.85$). There was also a significant difference for the simple interaction in the Stage IV manual ($\underline{M} = 18.76$, $\underline{SE} = 0.68$) condition compared to the Stage IV automated condition ($\underline{M} = 16.49$, $\underline{SE} = 0.92$) when pilots transitioned from a manual Stage III condition, $\underline{F}(1,7) = 32.69$, $\underline{p} < .05$.

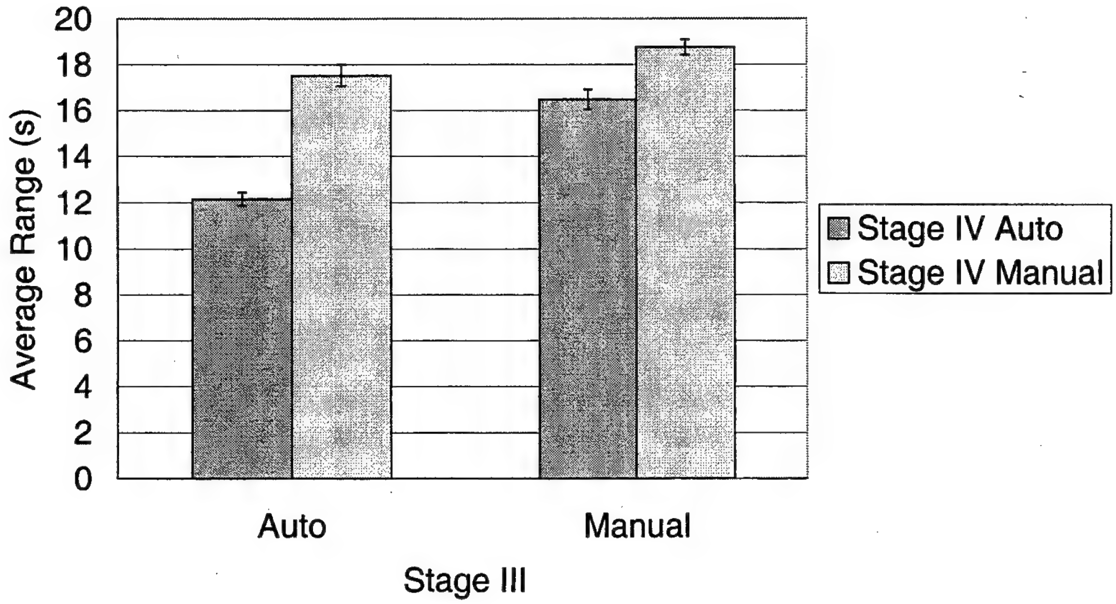


Figure 30. The Average-Range between shots (s) as a function of automation levels in Stage III and Stage IV. Error bars represent one standard error of the mean.

The Average-Range measure, as implemented here, has the potential to mask differences in the near and far groups. The Group 1 Range and Group 2 Range measures were implemented to capture any differential performance effects that may be present due to the serial order, and hence the amount of time after the Stage III transition, the pilot encountered the individual groups.

Stage IV Group 1 Range

The Group 1 Range measured the time needed for the pilot to get the shots off only for the first group encountered. The omnibus test revealed that there was a significant main effect for Stage III automation level, $\underline{F}(1,7) = 24.51$, $\underline{p} < .05$, and the two-way interaction between automation levels in Stage III and Stage IV, $\underline{F}(1,7) = 8.95$, $\underline{p} < .05$. When transitioning from an automated Stage III condition ($\underline{M} = 14.70$, $\underline{SE} = 0.53$) pilots were able to, on average, get the missiles fired at the targets 3.84s faster than when they transitioned from a manual Stage III condition ($\underline{M} = 18.54$, $\underline{SE} = 0.78$). The simple interaction comparisons of this interaction revealed that the significant source of the variance was the

difference between the Stage IV automated condition ($\underline{M} = 12.43$, $\underline{SE} = 0.64$) and the Stage IV manual condition ($\underline{M} = 16.79$, $\underline{SE} = 0.76$) after transitioning from an automated Stage III condition, $\underline{F}(1,7) = 23.57$, $\underline{p} < .025$. All other comparisons failed to reach a difference that was statistically significant ($\underline{p} > .0125$).

Stage IV Group 2 Range

The omnibus test of the Group 2 Range measure was not substantially different for the main effects of Stage III automation levels or the mission workload levels ($\underline{p} > .05$). It was, however, significantly different for the main effect of Stage IV automation level, $\underline{F}(1,7) = 44.20$, $\underline{p} < .05$. On average, pilots shot the targets in the far group 5.2s faster when Stage IV was automated ($\underline{M} = 13.36$, $\underline{SE} = 0.75$) as compared to the manual Stage IV condition ($\underline{M} = 18.56$, $\underline{SE} = 0.86$). All other tests lacked significance ($\underline{p} > .05$). This suggests that the influence of the Stage III automation level was not pervasive enough to make a difference in the amount of time the pilots needed to shoot the far group of targets (later in Stage IV) as it did in the near group of targets (early in Stage IV).

Global Measures

The global measures include the summed scores for the primary task ($P_I + P_{II} + P_{III} + P_{IV}$), secondary task ($S_I + S_{II} + S_{III}$), and Total scores ($T_I + T_{II} + T_{III} + T_{IV}$). Additionally, the subjective measures that were completed after each scenario are included. It was decided a priori that these measures would be submitted to a 2 × 2 (Stage I automation level × workload level) repeated measures ANOVA. This decision was compatible with the hypothesis that the Stage I automated and manual levels would have a sustained effect throughout the mission with regard to both performance and the subjective ratings of mental workload, situation awareness, trust, confidence and the rating of the automation's reliability.

Global Primary Task Scores

The omnibus ANOVA for the Global Primary Task Score revealed a significant main effect for Stage I automation level, $\underline{F}(1,7) = 38.46$, $\underline{p} < .05$, workload level, $\underline{F}(1,7) = 19.90$, $\underline{p}$

< .05, and the two-way interaction between these factors, $\underline{F}(1,7) = 12.47$, $\underline{p} < .05$. As demonstrated in Figure 31, pilots generally scored well on the primary tasks throughout the mission if they experienced a Stage I automated level, regardless of the workload level. The Global Primary Task scores were less when the pilots started Stage I in the manual condition and this effect was further mitigated by the workload level they experienced.

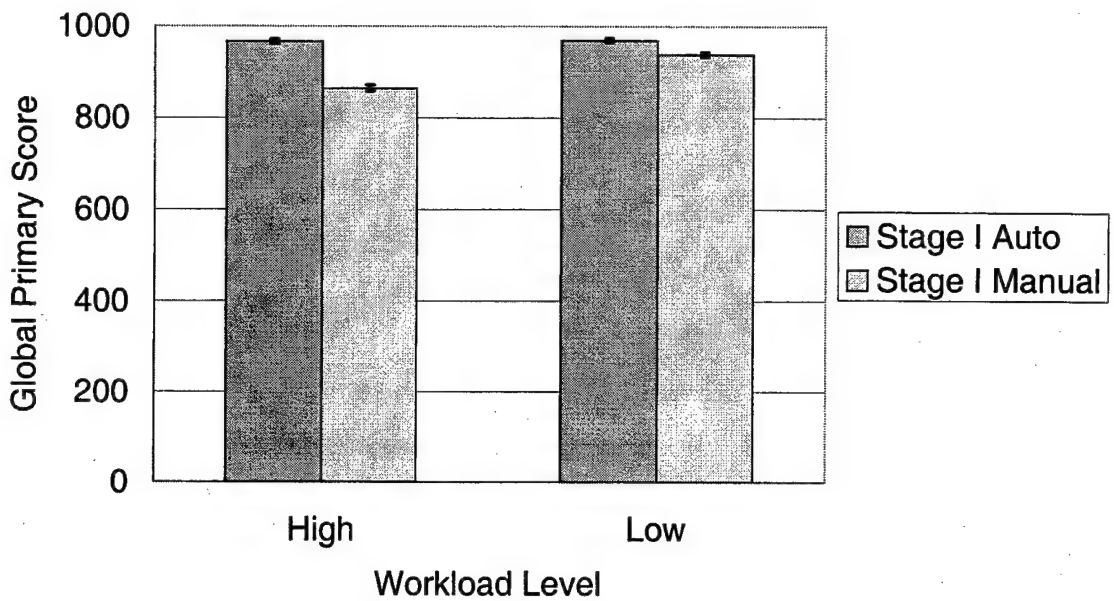


Figure 31. Total points scored on the primary tasks as a function of Stage I automation level and workload level. Error bars represent one standard error of the mean.

The post-hoc comparisons supported this interpretation. There was a significant difference in the Global Primary scores for the high workload condition, $\underline{F}(1,7) = 50.28$, $\underline{p} < .0125$, between the Stage I automated condition ($\underline{M} = 966.48$, $\underline{SE} = 9.97$) and the Stage I manual condition ($\underline{M} = 864.38$, $\underline{SE} = 12.82$). Additionally, there was a significant difference between the Stage I manual conditions, $\underline{F}(1,7) = 28.48$, $\underline{p} < .0125$, for the high workload condition and the low workload condition ($\underline{M} = 938.52$, $\underline{SE} = 9.59$).

Global Secondary Task Scores

The Global Secondary Task scores consisted of the sum of the Secondary Task scores for Stage I, Stage II and Stage III. The omnibus ANOVA conducted on these data revealed that there was not a significant main effect for either workload level or Stage I automation level ($p > .05$). The two-way interaction between these factors was similarly unremarkable ($p > .05$). This suggests that pilots, on average, were not affected by the workload manipulation or Stage I automation level when considering their ability to maintain their flight parameters for the entire mission.

Global Scores

The omnibus ANOVA for the Global scores revealed a significant difference for the main effect of Stage I automation level, $F(1,7) = 14.88$, $p < .05$, and the two-way interaction between Stage I automation level and workload level, $F(1,7) = 8.68$, $p < .05$. The distribution of the Global scores necessarily mirror (due to a lack of variance in the Global Secondary Task scores) that of the Global Primary Task scores (see Figure 31). The post-hoc comparisons conducted revealed that there was a significant difference in the Global scores for the high workload condition, $F(1,7) = 45.76$, $p < .0125$, between the Stage I automated condition ($M = 858.61$, $SE = 8.29$) and the Stage I manual condition ($M = 746.38$, $SE = 8.53$). Additionally, there was a significant difference between the Stage I manual conditions, $F(1,7) = 17.31$, $p < .0125$, for the high workload condition and the low workload condition ($M = 815.41$, $SE = 8.28$).

Subjective Measures

The subjective measures of mental workload, situation awareness, trust and confidence were collected at the end of each mission. These measures were analyzed using the same 2 × 2 (Stage I automation level × workload level) repeated-measures ANOVA. The method of collecting these data was previously outlined and the original forms are available for inspection at the end of Appendix A. For analysis purposes, each measure was reduced to

one data point for each scenario. The methodology utilized in the production of the one data point (where more than one was available) is listed in the appropriate sections below.

Mental Workload

The mental workload measure was determined by averaging across the six NASA-TLX sub-scales of mental demand, physical demand, temporal demand, performance, effort, and frustration. Each scale was rated from 1-100 with the left anchors indicating "low" and the right anchors indicating "high" for all scales except the performance scale, which is reversed. The average of the six sub-scales was computed after the pilot entered the six sub-scale values individually. This average has been found to be psychometrically equivalent to the weighted sub-scale averaging suggested by the NASA-TLX authors (Nygren, 1991). Empirically, the weighted averages have not been found to be superior to the simple average of the sub-scales (Christ et al., 1993; Hendy, Hamilton, & Landry, 1993). Additionally, there is some concern that the six subscales are often perceived as measuring only one or two constructs and interpretation of the individual subscales should only be made with caution (Bailey & Thompson, 2001). The omnibus ANOVA for the Average TLX scores revealed that there was a significant main effect for Stage I automation level, $\underline{F}(1,7) = 22.23$, $\underline{p} < .05$, and the main effect for workload level, $\underline{F}(1,7) = 13.25$, $\underline{p} < .05$. The two-way interaction did not significantly differ for these two factors ($\underline{p} > .05$). Figures 32 and 33 illustrate the main effect differences.

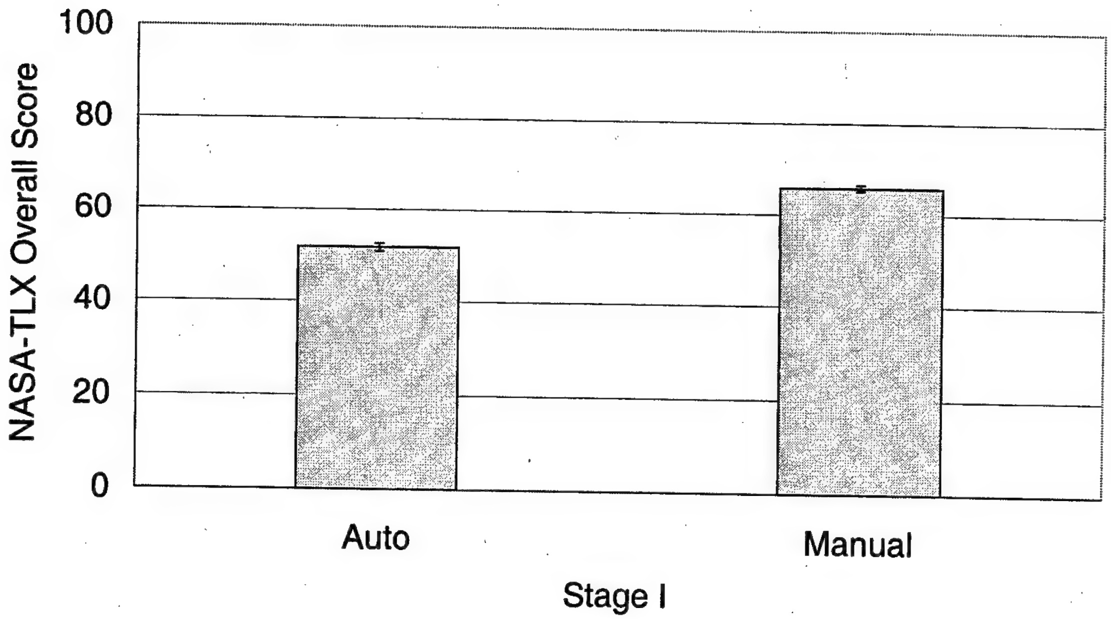


Figure 32. Average NASA-TLX scores as a function of Stage I automation level. Error bars represent one standard error of the mean.

As evident in Figure 32, pilots subjectively assessed their mental workload as being lower across the whole mission when they encountered automation in Stage I ($\underline{M}$ = 51.80, $\underline{SE}$ = 1.72) compared to when they performed Stage I manually ($\underline{M}$ = 65.89, $\underline{SE}$ = 1.29). This suggests that pilots based a significant amount of their subjective assessment of mental workload on the automation level experienced in Stage I even though they were instructed to make the assessment based on the entire mission.

The main effect for workload level was in the expected direction (see Figure 33). Pilots assessed their mental workload as being higher when the mission workload level was high ($\underline{M}$ = 60.38, $\underline{SE}$ = 1.64) as compared to when the mission workload level was low ($\underline{M}$ = 57.30, $\underline{SE}$ = 1.64). The difference between the levels of the workload factor (3.08) compared to the difference in the levels of the Stage I automation factor (14.09) suggests that the Stage I automation level was the principle factor that led to the differences in the subjective assessment of mental workload.

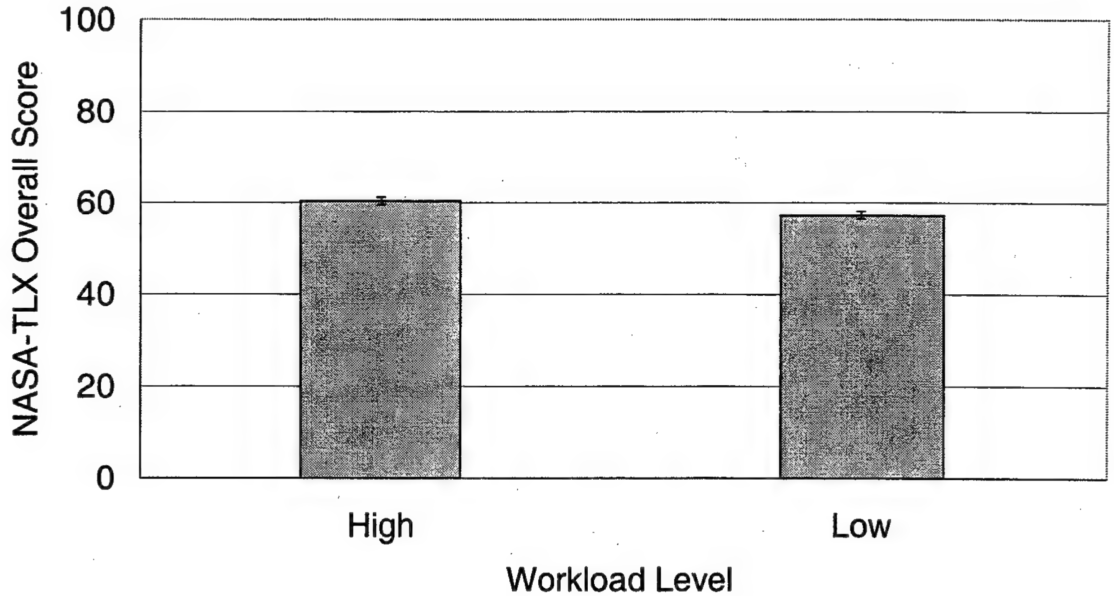


Figure 33. Average NASA-TLX scores as a function of mission workload level. Error bars represent one standard error of the mean.

Situation Awareness

The subjective measurements of situation awareness included the three items (see Appendix A) from the 3-D SART and the overall rating of situation awareness. Pilots were instructed to take into account the three SA sub-scales in their rating of overall SA. The overall SA rating was submitted to ANOVA analogous to that conducted for the mental workload scale. The omnibus test indicated that there was a significant main effect for Stage I automation level, $\underline{F}(1,7) = 5.70$, $\underline{p} < .05$, and a significant two-way interaction between Stage I automation level and workload level, $\underline{F}(1,7) = 5.83$, $\underline{p} < .05$). The two-way interaction, shown in Figure 34, illustrates that pilots perceived they had an overall better awareness of the situation (+0.88) when they started the mission off in an automated rather than manual Stage I condition. Further, their assessment of SA was decreased under high levels of workload compared to low levels of workload when they performed Stage I manually. The post-hoc comparisons supported this interpretation. There was a significant

difference in the ratings, F(1,7) = 30.58, p < .0125, across the high workload condition when the pilots had an automated Stage I condition (M = 5.75, SE = 0.13) compared to when the performed Stage I manually (M = 4.52, SE = 0.17). All other comparisons failed to reach a difference that was statistically significant (p > .0125).

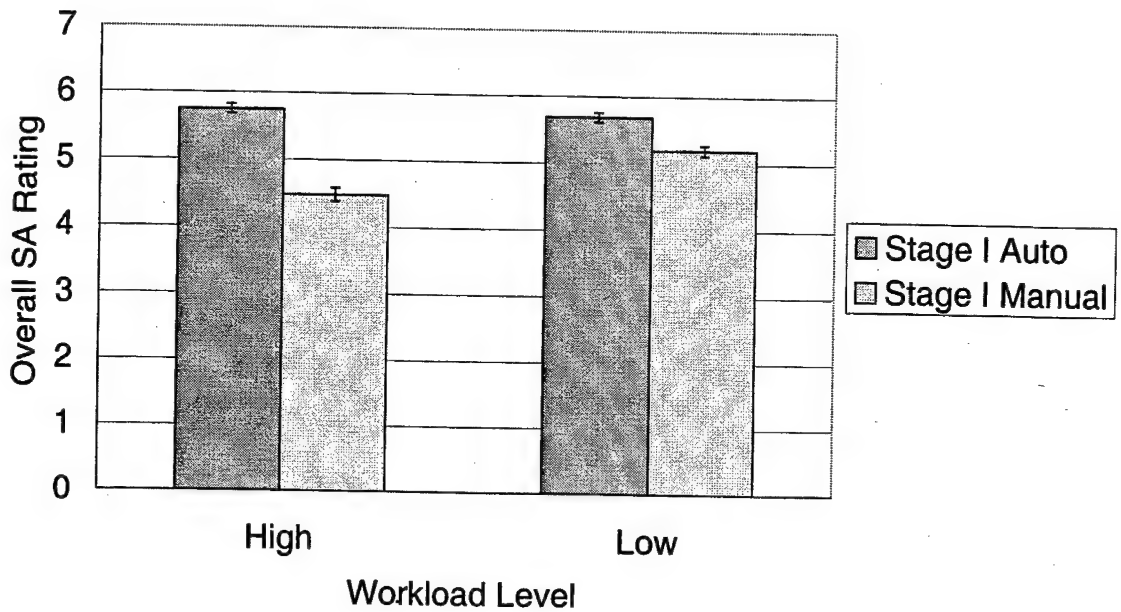


Figure 34. Overall SA rating as a function of Stage I automation level and workload level. Error bars represent one standard error of the mean.

Trust, Confidence and Reliability Ratings

The trust rating used in the analysis was the last question listed on the trust scale; "Overall, how much do you trust the system?" The omnibus test for the trust rating revealed the main effect of Stage I automation level was significantly different, F(1,7) = 5.91, p < .05, while the workload level main effect and the two-way interaction between these factors failed to reach a statistically significant difference (p > .05). A similar pattern of results was found for the Stage I main effects for the confidence the pilot had in completing their mission, F(1,7) = 8.62, p < .05, and the reliability rating of the automation, F(1,7) = 5.70, p <

.05. The mission workload factor did not have an appreciable effect on the pilot reported trust or confidence ratings ($p > .05$). Further, the mission workload level did not influence their perception of the reliability of the automation ($p > .05$). This indicates an average shift in the pilot ratings of the trust, confidence and reliability of the automation as a function of the Stage I automation level experienced. Table 8 lists the means and standard errors of the means (in parentheses) for the automated and manual Stage I levels for these three measures.

Table 8. Pilot trust, confidence, and reliability ratings by Stage I automation level. Standard errors of the means are in parentheses.

	Stage I	
	Automated	Manual
Trust	74.50 (1.30)	65.46 (1.57)
Confidence	75.37 (1.44)	64.26 (1.70)
Reliability	78.92 (1.36)	70.79 (1.75)

Overall, this data suggests that the Stage I automated condition increased the level of trust in the system, confidence that they could complete the mission, and increased the reliability rating of the automation. It should be pointed out that the reliability shift occurs with no change in the actual reliability level as the automation was perfectly reliable in this experiment.

Agreement of Subjective Measures

Three issues remain that need to be addressed with regard to the subjective measures. First, if the subjective measures are ranked by each type of scenario, is it useful to determine if there is agreement between each of rankings of the individual measures? Second, is it informative to look at the correlation between the subjective measures? Finally, how do the subjective measures compare to the performance data?

The preceding subjective measures were submitted to a Friedman chi-square test to determine if there was agreement in the ranked average ratings. There are two statistics that are produced when this test is performed. The first is the Friedman test statistic, which distributes as chi-square, which tests the null hypothesis that the rankings are not similar. This test was significantly different, $\underline{X}^2(4, N = 32) = 85.08$, $\underline{p} < .05$, indicating that there was substantial similarity between the rankings of the individual subjective ratings between the scenario types. In other words, a particular scenario type that was ranked high for overall situation awareness was similarly ranked high for overall trust level. While this is informative in determining the similarity or dissimilarity of the ratings it does not provide an exact indication of the degree of similarity between the measures. For this, the appropriate statistic is Kendall's coefficient of concordance ($\underline{W}$), which gives a numerical value between 0 and 1 based on the strength of the agreement between the measures. A value of 0 indicates that there is complete disagreement and a value of 1 indicates complete agreement. The coefficient of concordance, $\underline{W} = .665$, indicated a medium-high level of agreement between the ratings of the subjective measures tested.

A correlation matrix was produced (see Table 9) comparing the ratings for each of the subjective measures listed. The correlation coefficients generally demonstrated that a high or low rating on one scale consistently lead to a corresponding high or low rating on the other scales within the 32 different scenario types each pilot completed.

Table 9. Correlation coefficients between subjective measures within scenario types.

	TLX	SA	Trust	Confidence	Reliability
TLX	1.00				
SA	-.84	1.00			
Trust	-.78	.75	1.00		
Confidence	-.91	.85	.87	1.00	
Reliability	-.59	.52	.86	.72	1.00

The establishment of agreement however is not indicative of how the subjective measures compared to the performance as measured by the Global Score for each type of scenario. To determine the similarity between the Global Scores and the subjective measures, another Friedman test was conducted that included the Global Scores. The results indicated that there was a significant similarity between the subjective measures and the Global Scores collapsed across pilots, $\underline{X}^2$ (5, N = 32) = 129.34, $\underline{p} < .05$. The concordance measure ($\underline{W}$ = .808) indicated that there was a high degree of similarity between changes in the subjective measures and changes in the performance scores. Table 10 lists the associated correlation coefficients between the individual subjective measures and the Global Scores within scenario types.

Table 10. Correlation coefficients between subjective measures and Global Scores within each type of scenario.

	Global Score
TLX	-.77
SA	.85
Trust	.63
Confidence	.74
Reliability	.34

Table 11 lists the top eight scenarios ranked by Global Score. The associated experimental conditions are listed as well as the associated subjective measures for that scenario collapsed across pilots.

Table 11. Top eight Global Scores with experimental factors and associated subjective ratings (A= Automated, M = Manual).

Rank	Global Score	Workload	I	II	III	IV	TLX	SA	Trust	Conf.	Rel.
1	917.125	High	A	A	A	M	54.39	6.25	68.72	77.33	73.84
2	882.125	High	A	A	M	A	51.81	5.38	70.65	74.67	71.29
3	880.025	Low	A	A	A	M	57.79	5.38	72.57	70.63	80.06
4	880.125	Low	A	M	A	A	42.67	6.00	78.10	81.22	81.30
5	878.625	Low	A	A	M	A	46.74	5.75	75.15	78.26	77.80
6	872.250	Low	M	A	M	M	51.88	5.63	73.69	74.34	78.38
7	872.250	Low	A	M	A	M	58.91	5.25	71.29	67.08	74.64
8	865.000	Low	M	M	M	M	66.42	5.5	62.89	59.88	58.00

Introducing more automation is not always beneficial, both in terms of operator performance improvements and subjective assessment of the system. As Table 11 indicates, the conditions where all stages of the mission were automated did not reach the criterion set for the top eight scenarios as assessed by the Global Scores.

Post-Experimental Measures

De-Briefing Questionnaire

The post-experimental questionnaire, shown in Appendix D, was given to the pilots immediately following the data collection trials. The first question asked the pilots to indicate their agreement with the statement "Overall, the automation was helpful, when present, in completing the mission". On a scale of 1 (disagree greatly) to 10 (agree greatly) the pilots indicated they strongly agreed that the automation was helpful in completing the mission (<u>M</u> = 9.13, <u>SE</u> = 0.40). When asked to rank the stage the automation was most helpful, 8 out of 8 (100%) pilots ranked Stage I automation first and foremost. Seven of the pilots (87.5%) ranked Stage IV as the second most helpful out of the four stages. Stage II automation was listed as third most helpful by 7 of the 8 pilots (87.5%). One pilot had these

rankings switched; indicating Stage II was more helpful than Stage IV automation. All of the pilots (100%) agreed that Stage III automation was the least helpful when forced to rank all of the stages. These rankings were submitted to a Friedman chi-square test to determine the level of agreement between pilots. This test was significantly different, $\underline{X}^2 (4, N = 8) = 19.80$, $\underline{p} < .05$, indicating that there was a substantial similarity between the pilot rankings of the stage of automation that was most helpful. While, again, this is informative in determining the similarity or dissimilarity of the rankings it does not provide an exact indication of the degree of agreement between the measures. Kendall's coefficient of concordance , $\underline{W} = .825$, indicated a high level of agreement between the rankings of the most helpful stage of automation across all of the pilots.

The next question asked the pilots to indicate the best configuration for each stage of the mission if they were free to choose between manual and automated conditions. Concurrent with the answers to the last question, 100% of the pilots indicated they would prefer that Stage I was automated. Seven of the eight pilots indicated that Stage II should be automated, given the choice. It is noteworthy to point out that this question is decidedly different than the previous question. The result here indicates that given the opportunity to have automation, one of the eight pilots indicated that they would rather perform the tasks associated with that stage manually. The result is more profound in Stage III where only five of the eight pilots indicated that they would rather have an automated aid than perform the tasks manually. All of the pilots agreed that Stage IV should be automated given the discretion to choose.

The fourth question asked for pilot input on how much the automation improved their SA for each of the stages. A rating of "1" indicated that the automation "greatly decreased my SA" and a rating of "10" indicated that the automation "greatly improved my SA". The ratings, shown in Table 12, resemble the results of the rankings from the first question. That is, pilots indicated that Stage I automation was best at improving their SA and they also rated Stage I automation as the most helpful as addressed in the first question.

Table 12. Pilot ratings of improved SA with automation present by stage of the mission.

Stage	Mean	Standard Error
I	9.86	0.13
II	6.25	0.41
III	5.38	0.56
IV	7.75	0.65

The fifth and final question to be discussed asked the pilots to rate their reliance on the automation for each stage of the mission. A rating of "1" indicated they "did not rely at all on the automation" and a rating of "10" indicated they "relied greatly on the automation". The results for the reliance on the automation were similar to the previous and first questions. Stage I was relied on most ($\underline{M}$ = 9.23, $\underline{SE}$ = 0.26) followed by Stage IV ($\underline{M}$ = 8.00, $\underline{SE}$ = 0.50), Stage II ($\underline{M}$ = 5.88, $\underline{SE}$ = 0.88) and finally Stage III ($\underline{M}$ = 3.5, $\underline{SE}$ = 0.63).

In addition to the written questions, information was solicited from the pilots regarding their thoughts about the experiment, the strategies they employed under the experimental conditions, and general feedback regarding the experiment. Of particular interest was why they performed poorly on the secondary task in Stage I when the workload level was low and Stage I was automated. The pilots responded, in general, that they had the opportunity to explore more of the terrain using the automation during the low workload condition. While they were "playing" with the automation they did not pay close attention to the flight parameters and ended up deviating from the commanded flight path.

Post-Experimental Trial Results

As indicated in the Methods section for this experiment, each pilot was asked if they would volunteer to repeat four scenarios of their experiment. The purpose for this request was to determine the effectiveness of the training on the primary task and gauge, if possible, any learning effects across scenarios. Every pilot agreed to participate in the post-experimental trials without knowledge of which scenarios they would be given. The scenarios given to each pilot were the same as the first four scenarios, termed "pre" for this

analysis, that they completed in the data collection trials. The global points were submitted to a pre vs. post paired t-test which indicated that there was a significant increase, $t(31) = -4.38$, $p < .05$, in performance based on global points between the average pre trials ($M = 762.56$, $SE = 24.54$) and the post trials ($M = 860.16$, $SE = 18.27$). The 97.60 point difference was spread across all scenarios in the experimental trials due to the Latin-square design for starting scenario. Nevertheless, the difference was a concern due to the magnitude it represented and the lack of specific information it provided with respect to where the performance benefit was achieved.

To address this concern, a paired t-test was conducted for all Primary Task scores for each of the stages as well as the Total Secondary Task score. The results of this analysis, shown in Table 13, revealed that 66.50 points (68%) of the 97.59 point difference obtained for the Global score was due to an increase in the pilot's ability to fly the aircraft and maintain the secondary task parameters. The results further indicate that the only significant difference for primary task performance between the pre and post scenarios occurred in Stage III. This difference (14.06 points) represented the second largest percent gain (14%) of the factors that were included in Global scores. A noteworthy point regarding the Stage III difference is evident in the fact that all pilots achieved the maximum possible points in Stage III in the post-experimental scenarios.

Table 13. Average scores and differences for the repeated scenarios across pilots (* = p < .05).

Factor	Pre	Post	Difference
Global score	762.56	860.16	97.60*
Stage I Primary Task score	279.53	282.81	3.28
Stage II Primary Task score	148.75	155.63	6.88
Stage III Primary Task score	135.94	150.00	14.06*
Stage IV Primary Task score	295.00	301.88	6.88
Total Secondary Task score	146.66	80.16	66.50*

These results suggest that the pilots received adequate training in the primary tasks before starting the data collection trials. The robust training method is evident in the lack of significantly different scores in all but Stage III primary tasks ($p > .05$). Further, the secondary task difference, shown in the Total Secondary Task Score, indicates that a majority of the performance benefit between pre and post trials was achieved by more effectively maintaining the commanded flight parameters. This benefit however can be the result of either an increased level of competency in flying the simulator or it can be due to the ability of the pilot to effectively manage the dual tasks. If the benefit is spread across all stages equally, the argument is stronger that the pilot became more competent at flying the aircraft as the experiment progressed. Conversely, if the benefit is coupled with a particular stage, the argument is stronger that the pilot was better able to effectively manage the primary and secondary dual tasks. The results indicate that of the 66.50 Total Secondary Task score difference (the difference between the pre and post scenarios), 24.11% was attributable to a Stage I secondary task performance benefit in the post experimental trials. A 31.81% increase in secondary task performance was seen in Stage II and a 44.08% increase in Stage III. This distribution suggests that the performance benefit was not due solely to an increase in the pilot's ability to maintain the flight parameters throughout the experiment. Rather, it suggests that the pilot became more adept at attending to the secondary task as a function of the stage of the mission they were flying.

DISCUSSION

The purpose of conducting these studies was to determine what human-system performance differences exist as automation is implemented at different stages of the information-processing cycle. Furthermore, this group of studies examined the utility of applying the Parasuraman et al. (2000) model for types and levels of human interaction with automation. This section will compare the results of these studies to the results of previous studies reviewed in the introduction. The comparisons will be framed by the hypotheses made at the beginning of this thesis. In addition to the general purpose for conducting this group of studies there were objectives that were related to specific studies. For example, the reliability level of the automation, regardless of the stage it was in, was varied only in the set of studies that utilized the visual search task. Issues of this nature will be addressed as each study is considered.

Improving Performance with Automation

According to some authors, the introduction of automation may not always provide a concomitant increase in human-system performance. Rovira, McGarry et al. (2002) suggested that benefits of automation might be related to task complexity. If the manual task is relatively easy, automating it may not produce an increase in performance. Further, there is some evidence that performance improvements may be made only in difficult, high workload task environments (Merlo et al., 2000) or under temporally demanding or uncertain task situations (Dzindolet et al., 2001; Muthard & Wickens, 2001). The results of the visual search experiments support these observations. When the search task was relatively easy, under the 10 distractor set size, there was not a significant increase in the number of correct responses between the manual and automated conditions. Further, the automated cuing did not reduce the response times for correct responses. This supports the claim by Rovira, McGarry et al. that automation may not show a benefit when the task is relatively easy. This result was confirmed in the second and third visual search studies. In the low distractor set

size, the percentage of correct responses was not significantly different under the automated conditions than they were under the manual condition. The only exception, which occurred in both the second and third studies in the low distractor set size, was that the percent of correct responses was lower when the IA cue was coupled with the DA cue (together or separately).

The results of the visual search task experiments are not devoid of effects from the levels of reliability used in the experiments and those effects will be addressed below. The search and destroy study, however, did not vary the reliability level of the automation so a direct comparison can be made between the automated and manual conditions in that experiment. Remembering that we want to examine human-system performance, only the Total stage scores are considered here unless there is a compelling reason to include one of the other measures, which will be duly noted. The results indicate that the automation improved performance at every stage of the mission over manual performance with the exception of Stage III, where there was a slight, but not significant, drop in the Total Stage III score. These comparisons are based only on the Total score and the difference in the automated and manual condition within that stage. These results are in line with the first hypothesis; automation will facilitate better human-system performance when it is reliable.

There were, however, mediating factors that may offer a slightly modified interpretation. As stated in the results section in the search and destroy task experiment, the Total stage scores were analyzed by adding the automation levels in the previous stage (except Stage I) and the workload levels as additional factors. In stage I, the automation showed a clear benefit in the Total Stage I score. Under low workload conditions the difference was not significant between the automated and manual conditions. It was significantly different between these two conditions under the high workload condition. This result is consistent with the Merlo et al. (2000) suggestion that the benefits of automation may only reveal themselves under high workload conditions.

The Stage II scores did not show the same pattern of results. While the Total Stage II scores in this stage were higher under the automated condition compared to the manual condition the biggest difference in the Stage II score was due to the automation level in the

previous Stage I. This notion of automation transference between stages has not been examined previously. Prior studies have used performance measures based on the entire task that did not take into account performance that may have been stage specific. A similar pattern of results was noted for Stage III and Stage IV. The automation level in the previous stage also differentially affected the Total scores in the current stage and in the case of Stage III also by the workload level. This data supports hypothesis six in the Introduction section that stated that there would be a benefit in human-system performance in the current stage if preceded by an automated condition in the previous stage. Although transference has not been shown in this type of application it has been extensively documented in the examination of human error (Park, 1997; Reason, 1990), and accident investigations (Perrow, 1984).

The idea that there may be a feed-forward cost or benefit transference of performance will need to be rigorously tested and empirically validated in further research. In this case, most likely because the automation was perfectly reliable, there was a benefit that transferred. It is equally likely that a decrement could be transferred if the automation were not perfectly reliable. The level of transference may also be difficult to ascertain. The results, thus far, only indicate that there is transference between adjacent stages. This experiment also examined the Global score as a function of Stage I automation level. If the Global score can be affected by the automation level in Stage I, it can be assumed that the transference, at a minimum, is able to accumulate across all stages.

The data analysis of the Global scores indicates that this transference was possible and did occur. The Global score, a measure of overall mission effectiveness, was significantly different for the interaction between Stage I automation level and the workload level. Global scores averaged 77.67 points higher if Stage I was automated compared to when Stage I was performed manually. At this juncture it is appropriate to include a note of caution. It has been shown that a preceding stage that is automated, in general, has contributed to a higher performance level in the next stage. Also, the automation level affects the overall performance in the first stage. It would be easy to surmise that (a) the highest Global scores should come from trials that had automation in Stage I and (b) the highest Global score should come from the experimental condition where all of the stages

were automated. Table 11 showed the top eight experimental conditions ranked by Global Score. As stated in the results section, the experimental condition that had automation in all four stages was absent from this list. The reason for this is two-fold. First, the Global scores are also mediated by workload level, and second, the amount of transference between sets of stages is not necessarily equal. This issue will be revisited below when the discussion turns to mathematical modeling but the results do support the hypothesis that the automation level in the information acquisition stage significantly influences overall performance.

Effects of Workload Levels

The hypothesis regarding workload level changes states that the automation will show a greater effect on performance improvement under higher workload levels. The workload level in the visual search task was manipulated by increasing the number of distractors in the visual search area. In the first visual search study the workload manipulation involved either 10 or 20 distractors. The difference in the percent of correct responses between the manual and automated cueing was greater in the higher set size than for the lower set size indicating that the automation did help response accuracy when the search field was more saturated. In addition, the response times to the correct detections were faster in the automated cueing condition than the manual condition in the higher set size. This pattern was not evident in the smaller set size in which the response times were relatively equal. Furthermore, the percentage of trials that ended without a response was lower when the automation was present. The difference between the percent of timeouts was greater between the automated and manual conditions in the higher workload condition compared to the lower workload condition. In general, the automation in this study enabled more and faster correct responses and fewer non-response trials than the manual conditions in the higher workload conditions.

The results of the second and third study also support the observation that as the set size was increased the automation facilitated a higher percentage of correct responses. But this was true only for the information automation (IA) cueing condition. The decision-aiding (DA) cue, which recommended a course of action, did not show greater benefit under the

higher set sizes. The DA cueing condition was, in general, indistinguishable from the manual condition with regard to the percentage of correct responses as the set size was manipulated. Further, if the DA cue was combined with the IA cue, the percentage of correct responses dropped below the manual percentage of correct responses. This decrement was highest in the low workload condition and decreased as the set size was increased. This differential pattern of results was also seen in the third visual search study. The IA cue facilitated more correct responses and this benefit increased as the set size was increased. The DA cue again mirrored the results of the manual condition and the combined condition (IA + DA) remained relatively unchanged as the distractor set size increased.

These results suggest that there was a performance increase in the automation conditions under higher workload levels. But that performance increase was only observed for the IA cueing condition, automation that was present in the early stages of the information-processing cycle. Further, there was a greater cost of automation seen in the (IA + DA) condition. In general, the percent of correct responses remained the same across the increasing set sizes. This led to a decrement that was greater under the lower set sizes and decreased as the distractor set size increased. One possible explanation for this pattern of results is that participants became over-reliant on the automation in the combined condition. There is support for this interpretation if one can consider the combined automation cueing as a dual task. Parasuraman et al. (1993) found a complacency effect in a multi-task environment when the reliability level was either constantly low or high. Further, Molloy and Parasuraman (1996) did not find evidence of a complacency effect in a single-task environment. These results lend support for the results of the combined condition decrement. It is noteworthy to point out that the decrement gets smaller as workload levels increase because the other conditions are concurrently exhibiting a reduction in correct responses as the workload increases.

Another benefit of automation that was observed in the visual search studies was the reduction in the response times for correct responses as the set size increased. The IA cue provided this benefit in the conditions when it was presented alone or in the combined conditions. In all cases, if the IA cue was present there was a decrease in the response times

for correct responses over the manual and DA conditions. In both the second and third studies the reduction in response times increased as the distractor set size got larger.

In the search and destroy study the workload level also showed differential effects on performance between the automated and manual conditions. In Stage I, when the automation was present, the workload effect was negligible. In the Stage I manual condition, the performance was significantly degraded when the workload level was high compared to when the workload level was low. Thus, the automation nullified the workload effect in Stage I according to the performance measures. It would be expected that the workload level would have the greatest effect in Stage I because the workload manipulation directly made the primary task in Stage I more difficult by increasing the number of entities in each of the four groups.

There was not a significant difference in the ability of the automation to facilitate performance under either the high or low workload conditions in Stage II. In Stage III, the workload level did produce differential performance effects but the previous Stage II automation level mediated those effects. Under high workload conditions, the automation facilitated better performance but only if the previous Stage II was automated. If Stage II was performed manually, performance under the high workload conditions was reduced compared to the performance under low workload levels. This opposite trend was part of the reason why there was a significant three-way interaction between the automation levels of Stage II, Stage III, and the workload levels. There were no appreciable effects of the workload levels in Stage IV.

A final consideration regarding the ability of the automation to nullify the potential effects of workload manipulations is to consider the Global scores. There was a significant difference in the Global scores for the two workload levels but these scores were mediated by the automation level in Stage I. When Stage I was automated the Global scores were about equal under both levels of workload. If Stage I was performed manually however, there was a significant effect of workload on the Global score. So considering overall human-system performance, the data indicate that the automation, at least in Stage I, did nullify the workload effect as compared to when that stage was performed manually.

Effects of Reliability Levels

The effects of unreliable automation on human-system performance have dominated the bulk of the research thus far in studies that segment the automation into information-processing stages. Crocoll and Coury (1990) found that detection performance decreased when the unreliable decision aiding automation was present alone or in conjunction with the status or information automation. Similar results were also found by Rovira, McGarry et al. (2002), Rovira, Zinni et al. (2002), and Sarter and Schroeder (2001), and McGarry et al. (in press). The first visual search study had only one stage present and was represented by the IA cue. Even though only one stage was present there were differences noted between the automation that was perfectly reliable and the automation that was unreliable. For the percentage of correct responses there was a performance decrement between the reliable and unreliable conditions but only for the higher distractor set size. The data for the response times indicated that participants took longer to respond when they made a correct response when the IA cue was unreliable. For this measure, the response times were higher in the larger distractor set size than the smaller distractor set size. A similar pattern of results was obtained for the timeouts.

In the second visual search study, the percent of correct responses in the IA and DA conditions were both above the manual condition when the automation was reliable, as expected. When the automation was unreliable however, the percent of correct responses for both the IA and DA conditions fell below the manual baseline condition. This finding is not consistent with the results of previous studies when the magnitude of the decrement is evaluated. The difference in the IA condition was greater than the difference in the DA condition between reliable and unreliable automation conditions. In other words, unreliable IA cues in the information automation stage created a larger performance cost, in terms of the percentage of correct responses, than the unreliable DA cues in the decision-aiding stage. One can postulate that the reason for the inconsistent result is the nature of the task that was being performed. The visual search task was temporally compressed and a decision could not be made until either (a) the target was located, or (b) an exhaustive search was conducted

on the entire search field. In contrast, the Sarter and Schroeder (2001) task was based on a decision support system that emphasized the decision-making stage of the information-processing cycle. In addition, the duration of the flight task was much longer than that of the visual search task. The duration of the flight task was often in excess of 65s from the initial onset of the icing condition. The Rovira, McGarry et al. (2002) and McGarry et al. (in press) sensor-to-shooter task was also focused on decision-support. The trials were also longer (10s) than those in the visual search task. It can be argued that the visual search task is more of a perception task than a decision-making or decision support task. It may be the case that the effects of unreliable automation are task dependent. In higher order, more cognitively demanding tasks, the unreliable automation may have a more detrimental effect in the decision-aiding stage while in lower cognitively demanding tasks the detrimental effect may be tied to the earlier information automation stages. Wickens and Carswell (1997) provide a plausible explanation for the differing decremental effects. They posit that the number of transformations to the raw data that the human needs to make will increase the time and complexity of the overall information-processing cycle. This notion will have to be empirically tested by including task complexity as an experimental factor.

The results of the third visual search study were also informative with regard to the reliability level of the automation. In terms of the percentage of correct responses, the IA cue consistently lead to higher performance over the manual condition, regardless of the reliability level of the automation (50%, 70%, or 90%). The DA cueing condition however only surpassed the manual condition when the automation was at the 90% reliability level. Otherwise, the DA conditions were about the same (70% condition) or lower (50% condition) than the manual condition for the percentage of correct responses. Additionally, performance was consistently lower for the DA cueing condition than for the IA cueing condition. This data suggests that there was a performance decrement in the decision-aiding stage for correct detections as compared to the information automation stage. The DA condition performance did not however go below the manual performance until the level of the automation reliability was chance.

The response times to correct responses also revealed a differential effect for the level of reliability by the stage the automation was employed. For the DA cued condition, the response times were consistently close to the response times in the manual condition across all automation reliability levels. The IA cued conditions demonstrated a performance improvement over the manual condition and the DA cued condition as the reliability level of the automation increased.

Unreliable automation did have an effect on the performance of the participants. The third hypothesis that stated the cost would be greater in the decision-aided stage, met with mixed results. The second study did not support that claim. The percentage of correct responses showed a performance decrement in both the IA and DA conditions over the manual condition when the automation was unreliable compared to when it was perfectly reliable. The third study did support the hypothesis in that the unreliable DA cued condition consistently lead to poorer performance than the unreliable IA cued condition. But performance on the unreliable DA cued condition did not fall below the manual performance until the automation was at a 50% reliability level. This result supports the fourth hypothesis in that the lowest reliability group experienced the greatest performance decrement. With mixed results like this it would be premature to speculate what the exact nature of the relationship is between performance costs and the stage the unreliable automation was present. In light of the fact that the results obtained in the present studies were not consistent with previous results indicates that this is a good candidate for further experimental exploration.

Evaluation of the Model

The fifth hypothesis suggested that the Parasuraman et al. (2000) model would be scalable to other domains and to more complex environments. The measure of the scalability was proposed to include the amount of useful information that was obtained by using the model. This information is required to be over and above the information that would be available without using the model. Specifically, because the model is stage based, it should

provide information about human-system performance that is stage specific. It should also provide this information regardless of the domain area, task complexity, and scope of the task or tasks to which it is applied. The model should also, in experimental settings, be able to help in the design of experiments and the interpretation of the results.

By all of the standards listed, the model should be considered scalable, at least in its application to the experiments described here. The model served to design the visual search tasks and the automated cues that were of experimental interest. The model served as a suitable framework (and ahead of any previous models) to analyze the data and to interpret the results. Although not all of the results from the visual search task were completely consistent with the results obtained in previous studies, the framework the model provided allowed the difference to be analyzed systematically.

The application of the model to the more complex search and destroy task was equally favorable. By applying the model in the early experimental design process, specific stages of the overall task were identified and segmented. This allowed for the identification of potential tasks within the stages that were suitable to automation level manipulations. Once those primary tasks were identified, the differences in the manual and automation levels were instituted. The secondary tasks were added in a similar manner. The model helped to identify stage transitions in the overall task and those transition points became waypoints in the flight. The waypoint segments served as transition points for the secondary tasks. For example, if an altitude change occurred, it occurred at one of the transition waypoints. The model was able to be applied in this complex task (as compared to the visual search task for instance) environment that was in a different domain than it had previously been applied.

In addition to providing help in the design of the experiment the model also served as the basis for structuring the data analysis. As was apparent from the results of experiment four, the analysis did provide useful information about the performance differences in the primary and secondary tasks and performance as it related to meeting the overall mission objectives. Further, the structure of the model and hence the experiment, permitted the

analysis of performance differences between stages, something that would not have been available if that design structure had been absent.

The model can however be improved upon. As Miller and Parasuraman (in press) point out, there are improvements that can be made to the model architecture. They posit that the model does a decent job if one wants to consider the aggregate parent task only. They point out that a parent task may, and usually is by definition, composed of layers of sub-tasks that can, and possibly should, function at a level of automation that may be different than the level that was assigned to the parent task. They argue that the model, as it stands, does not decompose the parent tasks into sub-tasks that are functionally relevant. They propose that other decomposition methodologies be applied to the parent task and that those sub-tasks should be evaluated for the appropriate levels of automation, exercising this recursively until the sub-tasks are considered at their primary level. The result at the end of the exercise will be the application of appropriate levels of automation to all of the sub-tasks. This in turn will lead to better management of the roles the human and the automation will take part in together or separately.

Future Research

It should be clear from the previous pages that the experiment utilizing the SIRE facility was complex not only from a task perspective but also from a software development perspective. The complexity however did not stop there; there were hardware issues, training issues, scheduling issues, and time and budget constraints. It seems obvious that the next step in this research effort is to evaluate the effects of unreliable automation by stage utilizing the same search and destroy mission paradigm. These plans are underway and data collection on that study will commence shortly.

Outside of the SIRE facility, there are several research avenues that could be explored. For example, the relationship between task complexity and the effects of unreliable automation by stage should be examined to determine if there is a difference between more or less cognitively demanding tasks and the stage that exhibits the greater

decrement on performance. It is also important to consider task duration as a potential contributing factor in determining what stage may adversely affect performance as automation reliability is manipulated. There should be some further exploration into the notion that a performance transference, either positive or negative, may occur between or across stages in this framework.

To address this notion, there has been some effort to quantify the results of the search and destroy mission. The quantification takes place in the form of an equation that is currently being tested. It is important to note that the values used were determined by the point structure of the search and destroy mission, the values can change for different tasks and different performance metrics. The idea is relatively simple; in the case of determining if automation produces better scores the following equation would be used;

$$P = T_{I(A\text{-}M)} + T_{II(A\text{-}M)}\,(R_{(IA)}) + T_{III(A\text{-}M)}(R_{(IA)(IIA)} + T_{IV(A\text{-}M)}\,R_{(IA)(IIA)(IIIA)}$$

Where P equals the predicted score and would be equal to the Total score difference between the automated and manual Stage I conditions plus the Total score difference between the automated and manual condition in Stage II times the residual value of having Stage I automated, plus the Total score difference between the automated and manual condition in Stage III times the residual value of having Stage I and Stage II automated plus the Total score difference between the automated and manual condition in Stage IV times the residual value of having Stage I, II, and III automated.

The important thing to note is that the Total scores are additive and stage specific while the residual values are cumulative and multiplicative. Also, the residual value can either be positive or negative so that either a benefit or cost will transfer between stages. Of course this equation will need additional data and refinement but it is a start to the mathematical modeling that may be of value as the cost of testing real systems rises to prohibitive levels.

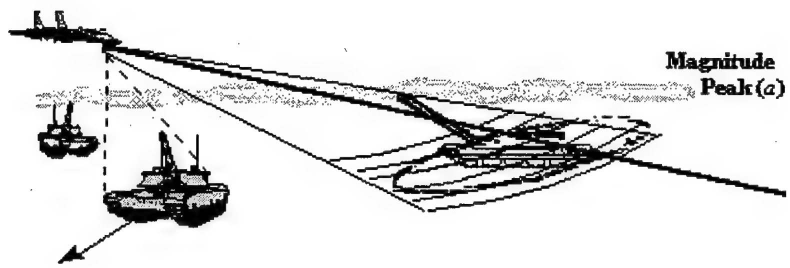
The modeling effort is ongoing and will take some time due to the fact that there are not a substantive number of experiments that have been conducted in this task environment.

In order to mature, the model will also need to take into account the effects of workload, situation specific task anomalies and other factors that have yet to be revealed. The key goal of this research effort was to advance the understanding of the relationship between humans and the automated systems that they utilize. By applying the prevailing methodologies to new task environments some positive steps have been made in that direction.

APPENDICES

APPENDIX A

AUTOMATION STUDY TRAINING MANUAL



Synthesized Immersion Research Environment (SIRE)
Air Force Research Laboratory
Wright-Patterson Air Force Base

Points of Contact:

Principle Investigator: Scott Galster, AFRL (937) 255-8737
SIRE: Matt Middendorf, MSSI (937) 255-0895
SIRE Dome: WPAFB, Area B, Bldg 33, Room 1001 (937) 255-0895
Pilot Pool: Becky Brown, Sytronics (937) 255-0884, DSN 785-0884

It is important that participants carefully read this training manual before starting the experiment

This document provides potential study participants with the following information:
Facility – An overview of the SIRE facility
Introduction – A description of the purpose and goals of the study
Design – Preliminary instructions to participants
Scenario – Description of close air support mission

1 Table of Contents

2 List of Acronyms

A/G	Air-to-ground
AD	Attack Display
APC	Armored Personnel Carrier
CMS	Countermeasures Management Switch
DD	Defensive Display
DMS	Display Management Switch
FEBA	Forward Edge of the Battle Area
FOR	Field-of-regard
FOV	Field-of-view
HDD	Head-Down Display
HOTAS	Hands On Throttle And Stick
HUD	Head-Up Display
RBGM	Real Beam Ground Map
RWR	Radar Warning Receiver
SA	Surface to Air defense system
SAM	Surface-to-Air Missile
SAR	Synthetic Aperture Radar
SD	Situation Display
SIRE	Synthesized Immersion Research Environment
TMS	Target Management Switch

3 List of Figures

4 List of Tables

5 General Background and Description

The Synthesized Immersion Research Environment (SIRE) is a state-of-the-art virtual environment research facility whose mission is to develop and evaluate advanced, multi-sensory virtual interfaces for future United States Air Force crew stations. The facility consists of several autonomous research stations that can support individual research efforts or be combined to form a multi-participant virtual environment. The primary SIRE cockpit is described in this document.

The SIRE cockpit includes a simulated F-16 shell and is fitted with an F-16C throttle and a side-mounted control stick. This is a fixed-base simulator situated in front of a 40-foot diameter dome that includes a high-resolution, large field-of-view interactive visual display. All system controls are accessible on the stick and throttle. Most controls not relevant to the task are automated or eliminated. The simulation is controlled from computers located in an adjacent area. Computing power for the simulation and all the displays is provided by a number of personal computers. Cockpit displays available are a head-up display (HUD), three head-down displays (HDD).

6 Introduction

The introduction of automation into highly complex systems has occurred under several guiding principles. The application of these principles has often resulted in tenuous interactions with regard to human performance within complex systems. With advances in technology increasing at an exponential rate it is no longer applicable to look at single automated tools but rather at how several automated tools fit together and affect system performance.

The current project will focus on three main objectives: Is there a model that takes into account the automated tools function within a complex task environment and can that complex task environment be broken down into functional areas that correspond and conform to information processing and decision-making processes that we find in human cognition? The second and third objectives will test the hypothesis that there is a functional congruency of complex system performance to that which is seen in human cognition. These objectives will be tested in the Synthesized Immersion Research Environment (SIRE) located at the Wright Patterson Air Force Base in Dayton, Ohio.

The following manual will give you an overview of the experiment you will participate in and some information regarding the tasks you will be required to perform.

7 The Mission

7.1 Layout

The scenarios used will be performed in the Synthesized Immersion Research Environment (SIRE) over a terrain database that extends from 109 degrees West to 111 degrees West and from 31 degrees North to 33 degrees North. This area is centered approximately 60 nm east of Tucson, Arizona. US overhead assets have sensed the movement of a large number of adversary vehicles near Benson, AZ. The forward edge of the battle area (FEBA) has been placed to encompass the adversary positions. The adversary vehicles are grouped together in four distinct areas (triangles) and are engaged in entrenching maneuvers (see Figure 1).

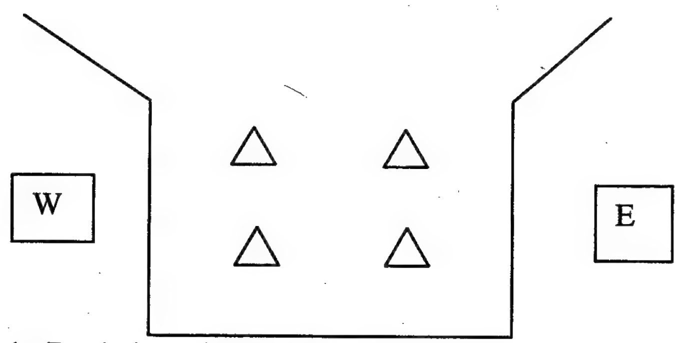


Figure 1: Depiction of the battle area.

7.2 Approach

There will be two mission-starting locations utilized, one from the West and one from the East. Each of these starting locations is associated with a prescribed route to the battle area. The starting points are located approximately 63 nm from the center of the battle area to the East/West.

7.3 Target Selection

You will be responsible for adding a total of eight targets to a shootlist that will be located on the HUD. Four of these targets will be from one of the groups nearest you and four will be from one of the groups furthest away from you as you approach the battle area. Specifics of how to add these targets will be addressed in later sections.

7.4 Course

You will be responsible for adhering to a commanded flight path, altitude, and airspeed until you reach the FEBA. While approaching the battle area you will be required to select a tactical flight path that best represents the targets you have chosen. An explanation on how to perform this task is also discussed in a later section

7.5 Time Constraints

Each mission lasts for a maximum of 10 minutes or until you cross the opposite side of the FEBA, whichever occurs first.

7.6 Threats

The SIRE is an unclassified flight simulation. Specifications of the simulated threats used in this study do not necessarily represent actual capabilities. The simulated threats are listed in Table 1.

Vehicle Class	Force	Identity
Tank	Friendly	M1A1
	Hostile	T-72
Armored Personnel Carrier	Hostile	BMP-2
Surface to Air Defense Systems	Hostile	SA-6 FCR, SA-6 TEL, SA-15, SA-9, ZSU-23

Table 1: Vehicles in and around the battle area.

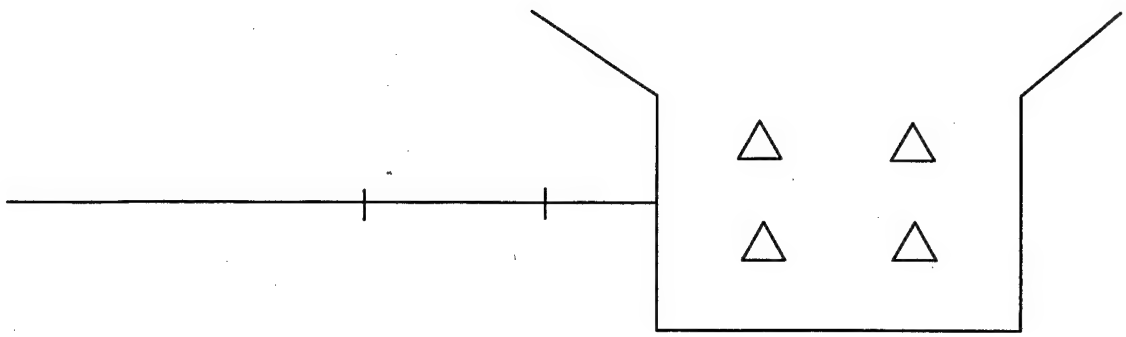
7.7 Environment

MODSAF 5.0, unmodified for this study, creates the actions of the adversary vehicles, air defense threats, and friendly vehicles. The MODSAF 5.0 actions are embodied in this simulation using a Distributed Interactive Simulation (DIS) protocol.

8 Stages of the mission

8.1 Overview

Each mission is segmented into four distinct stages. The stages are: 1) Target selection, 2) Target prioritization, 3) Course selection through the battle area, and 4) Weapon delivery. Figure 2 depicts the waypoint/stage transitions that correspond to the stages outlined above. The objective is to complete all the necessary tasks assigned to the stage before reaching the next stage while maintaining course, airspeed and altitude requirements. Table 2 lists all the relevant airspeed, altitude and distance information for each stage. You will start each mission on the proper heading, at10,000 ft. with an airspeed of 338 kts.



STAGE **I** **II** **III** **IV**

Figure 2: Depiction of a typical scenario.

Each of the four stages will be done in a manual mode or with the aid of automation. The level of automation for each stage will be displayed to the pilot at all times. Below, each stage will be described for both the manual and the automated modes. The tasks that need to be completed for each stage are also described. The transition between stages is waypoint driven. Distance and time to go between waypoints will be displayed. A transition between stages is signaled by an auditory alarm about ten seconds before the actual switch. This serves to alert the pilot to complete all necessary tasks that are outstanding and to prepare for the tasks in the next stage.

STAGE	DISTANCE TO FEBA	COMMANDED ALTITUDE	COMMANDED AIRSPEED
I	**37.5 nm**	10,000 ft.	450 kts.
II	15 nm	12,000 ft.	500 kts.
III	5 nm	14,000ft.	500 kts.
IV	AT FEBA	14,000ft	600 kts.

Table 2. Airspeed, altitude and distance from the FEBA for each stage.

* NOTE* Switchology for the functions described below will be addressed in the static training trial and in Section 6.6.

Before describing the specifics of each stage, it will be important to discuss the shoot list and its functionality, as it will be used throughout the mission. The shoot list contains eight slots that will be filled with the targets that you will select in stage one. The shoot list can function in several modes; ADD, DEPLOY, INSERT, REPLACE, MOVE, and OFF. Some of these modes only work when in a particular stage of the mission. Table 3 lists what modes are available in each stage of the mission. You should refer back to this table as needed as you read through the remaining descriptions.

		Shootlist Mode						
		ADD	DEPLOY	INSERT	REPLACE	REMOVE	MOVE	OFF
Stage	I	X	X	X	X	X		X
	II		X				X	X
	III		X					
	IV		X					

Table 3. Shootlist modes available for each stage.

8.2 Stage One – The Task

The task in stage one is to select the proper targets and add them to a shoot list displayed on the HUD. There are eight slots on the shoot list. The layout of all the trials is such that there are two near groups and two far groups of targets. The first step is to evaluate which group in the near set has the overall highest priority. The type of targets in the two groups will determine this. Table 4 lists the priority of the targets from SA-6 FCR as the highest to T-72 the lowest. This list will be available in the cockpit for reference.

Once the highest priority group has been identified, the highest priority targets **in that group** will be put on the shoot list in slots 1-4. This task will then be repeated for the two far groups and the selected targets will be put in slots 5-8 on the shootlist. Targets should only be chosen from one group for each set.

Each trial begins with the aircraft at 10,000 ft traveling at 338 kts at a heading of 90 or 270 degrees depending on the starting point. Once the trial starts, you will be required to adjust your airspeed to comply with the commanded airspeed and begin searching for targets. The Real Beam Ground Map (RBGM) radar is the default setting for the radar display. You will increase the range of the radar until entities appear on the radar screen. Entities are indicated on the radar display with a small box. The box may have a label above it indicating its current level of classification.

A cursor will be used to identify an area of interest on the radar screen. The cursor map can be used in a snowplow, ground stabilized, or locked mode. A Synthetic Aperture

Radar (SAR) patch map can then be made for the area of interest. When the SAR patch map is first made the ground stabilized mode is the default setting. To this point, there are no differences in the manual and automation enhanced manipulations.

8.2.1 Stage One – Manual

In the manual mode, the entities in the SAR patch map are subjected to a classification algorithm based on several factors such as size, shape etc. The classification of each target is promoted as a function of the amount of time the classifier is processing that unit (time the unit is in the SAR patch map). A typical promotion grid is listed in Table 4. As a unit or group of units promote, they step from a no label return to unknown through force then class and finally identity. The radar will display the current promotion step for each unit.

	Classification Level				
Entity	**No Label**	**Unknown**	**Force**	**Class**	**Identity**
SA-6 FCR	White Box	"UNK" (White)	"Enemy" (Red)	"SA"	"SA6 FCR"
SA-6 TEL	White Box	"UNK" (White)	"Enemy" (Red)	"SA"	"SA6 TEL"
SA-15	White Box	"UNK" (White)	"Enemy" (Red)	"SA"	"SA-15"
SA-9	White Box	"UNK" (White)	"Enemy" (Red)	"SA"	"SA-9"
ZSU-23	White Box	"UNK" (White)	"Enemy" (Red)	"AA"	"ZSU-23"
BMP2	White Box	"UNK" (White)	"Enemy" (Red)	"APC"	"BMP2"
T-72	White Box	"UNK" (White)	"Enemy" (Red)	"TNK"	"T-72"
M1A1	White Box	"UNK" (White)	Friendly (Blue)	"TNK"	"M1A1"
M113	White Box	"UNK" (White)	Friendly (Blue)	"APC"	"M113"

Table 4. Promotion steps for identifying mission vehicles. Vehicles are listed in prioritized order.

You will use the cursor to move around the area of interest and decide if this is the correct group for this set. If you decide this is the correct group, move the cursor close to the target, lock-on to it, and add it to the shootlist. Entities are added sequentially to the shoot list in this stage. This process will continue until 4 targets have been added from one of the near groups and 4 targets from one of the far groups. If you want to explore another group, you can either revert to RBGM and remake a SAR patch map or slew the patch map over to

the other group using the cursor while using the Situation Display for reference. Targets inadvertently added to the shootlist can be removed.

(Note: There will be ample training on how the switchology works and the data trials will not proceed until you have indicated that you have mastered the functions)

Once the shootlist is complete you will press the confirm button (missile step) signaling that you are finished with the stage one task.

8.2.2 Stage One – Automated

The automated aid in this stage incorporates information from various sources and identifies all entities prior to starting the mission. Thus, when you make a SAR patch map, the entities will already be promoted to their highest level. You still need to choose a group from the first set and add the highest priority targets in slots 1-4 then choose a group from the far set and add the highest priority targets in slots 5-8. Once this task is done, you will need to press the confirm button.

The ADD function only works in stage one of the mission; you cannot add more targets once the transition to stage two has occurred.

8.3 Stage Two – The Task

Stage two is concerned with prioritizing the shootlist according to location (the closest group first) and threat to ownship. The prioritized list should have four entities from the closest group ordered by threat (see Table 4) and four entities from the far group ordered in the same manner. It is important when prioritizing the list not to interchange targets between the first four slots on the shootlist and the last four slots, as this will complicate the stage three course selection and stage four weapon delivery.

8.3.1 Stage Two – Manual

In the stage two manual condition, you must use the MOVE mode on the shootlist to move targets up and down the list. This is done using switchology similar to "drag and drop" functions on a personal computer. Once the list is organized to your satisfaction the confirm button should be pressed.

8.3.2 Stage Two - Automated

In the automated condition, the computer will prioritize the shootlist automatically. You will be able to follow the steps as the computer is prioritizing the list. After the computer has finished making changes, if any, you must press the confirm button to accept the changes. You have the option of making changes to the list if the computer's list is unacceptable. If this is the case, make the changes and then press the confirm button.

8.4 Stage Three – The Task

Stage three is where you choose the tactical flight path to be flown through the battle area based on the groups of targets that were chosen. There are four groups (Near/North, Near/South, Far/North, and Far/South) hence there are four possible routes to choose from for each approach direction. You should choose the path that matches the targets on the shootlist whether they are the correct targets or not. The tactical flight paths will be shown on the Situation Display as connected waypoints through the battle area (see Figure 3). The first segment of the path will lead toward the near group, the next segment leads back to a central point and towards the far group chosen. The final segment leads to the outbound side of the FEBA.. The target area the flight path leads to will be highlighted in red on the Situation Display and will change as you cycle through the four potential courses. Figure 2 depicts the chosen target areas with circles inside the triangles. Regardless of the automation level in this stage, the SD will display one of the four flight paths randomly at the beginning of the stage.

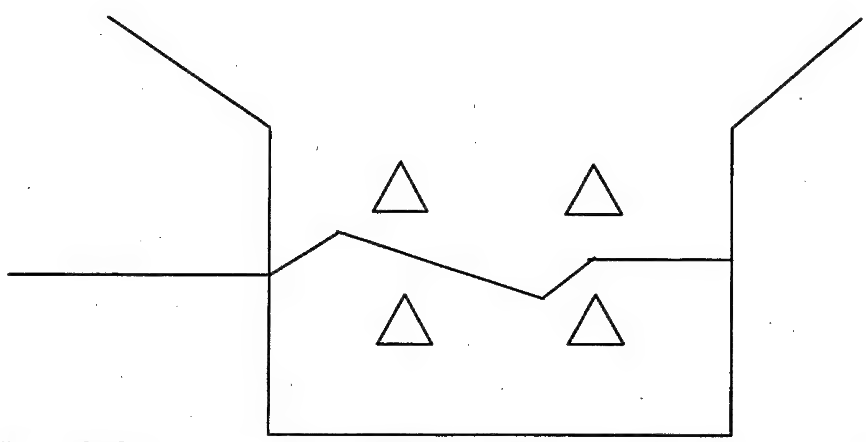


Figure 3. One possible tactical flight path through the battle area for a West approach.

8.4.1 Stage Three – Manual

You will toggle the pinkie switch until the best course is shown on the Situation Display. You will then press the confirm button signaling acceptance of the chosen flight path.

8.4.2 Stage Three – Automated

The computer will cycle to the best path for the targets that are present in the battle area. If you have chosen targets from a different group, the pinkie switch must be used to

choose the path that reflects the targets that are on the shootlist. The confirm button must be pressed to accept the flight path.

8.5 Stage Four

The final stage begins when the aircraft crosses the FEBA. Once inside the FEBA, the major tasks are to fire the missiles at the targets that are on the shootlist and avoid being hit by any adversarial munitions. **If fired upon, maintenance of altitude, airspeed and flight path become secondary, avoid being hit first then resume adherence to the flight parameters.** The course selected in stage three should lead you towards the chosen group in the near set. The first target on the shootlist will be locked-on to upon entry into stage four. Table 5 defines the parameters needed to provide a good firing solution on the targets.

Parameter	Value
Target	Designated
Range	< 12nm
Roll	+/- 30 deg
Pitch	+/- 20 deg
Gimbal Limits (missile)	30 deg
Altitude	< 14,800 ft (MSL)

Table 5. Requirements for missile lock.

8.5.1 Stage Four – Manual

In the manual mode, you must ensure that the shooting parameters (range, roll, pitch etc.) are within proper value by checking your instruments. A target designator box (TD box) will be displayed on the HUD showing the location of the target in the out-the-window scene. If the target is located outside the HUD FOV a locator line will appear with the degrees to target displayed to the left of the gun cross. It is important to note that the TD box may not be on the HUD even though the target is within the shooting parameters (i.e. gimbal limits).

After launching the first missile, you must manually designate subsequent targets. This is done by unlocking the current target, scrolling down the shootlist, and locking-on to that target. For maximum success, you must ensure the proper shooting parameters for each missile launch.

8.5.2 Stage Four – Automated

In the automated mode, launch codes will appear below the shootlist to advise you on the shooting parameters. The possible codes are; UNDES, RANGE, GIMBAL, PITCH, ALTITUDE, AND ROLL. If one of these codes appears below the shootlist you will know what parameter needs attention before a missile is fired. The messages are in priority order as listed above, that is, a roll message would not appear if the target is outside gimbal limits. All proceeding parameters must be met before the next code is considered and displayed.

The automation in this stage also provides for relief in the designation of subsequent targets. Once a missile has been fired, the automation automatically de-selects the highlighted target, scrolls down one slot, and locks on to the next target.

After all missiles have been fired, the pilot will proceed to the exit waypoint located on the FEBA line. The trial is over when the pilot crosses the FEBA line or the expiration of 10 minutes, whichever occurs first.

9 Experimental Design

After the completion of all applicable forms (biographical information, consent forms, pre-simulator sickness ratings) you will be able to get in the cockpit and fly around to get the feel of the simulator flight dynamics and controls. A training session (1 static and 8 dynamic trials) will be completed to familiarize you with the cockpit displays and switches, the mission, the stages of each mission, and the levels of automation for each stage.

Next, data collection sessions will be conducted in blocks of eight trials (missions). After each trial, you will answer subjective questions relating to the trial you just completed. These subjective measures will include the NASA Task Load Index (TLX), the Situation Awareness Rating Technique (SART), and a trust and confidence scale (see Section 7 for paper versions of these scales).

After the completion of all data collection, you will be asked to fill out a debriefing questionnaire relating to the experiment. It is anticipated that you can complete test requirements in (3) four-hour sessions. This is also the preferred test schedule, although an attempt will be made to accommodate alternate test schedules. A pre and post simulator sickness rating form **must** be filled out for each session.

10 Information on Simulator Displays/Controls

10.1 Overview

Dedicated cockpit displays provide situation awareness and offensive and defensive system control in the SIRE cockpit. Figure 4 shows the display arrangement. The left head-down display (HDD) provides defensive system control. The right HDD provides situation awareness. The center HDD is a display magnified at eye-level for wide field-of-view and

provides the interface for air-to-ground (A/G) sensors (RBGM and SAR). The head-up display (HUD) provides flight and status information. The HUD symbology adheres to MIL-STD-1787 with minor modifications.

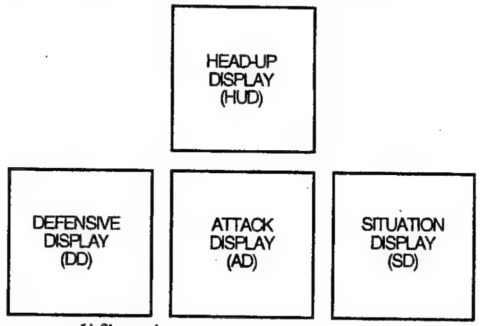


Figure 4. Display arrangement for the SIRE cockpit.

10.2 Situation Display (SD)

The SD presents data that aids in general situation awareness. The SD could be considered a dynamic version of a paper mission chart. The situation display (SD) shows the SIRE ownship (o/s) aircraft, target locations, and provides navigation information to enhance situation awareness.

The SD display shows an o/s symbol surrounded by four range rings. The innermost range ring includes a compass rose with cardinal headings (N, S, E, and W) and index marks every 30°. This inner ring rotates for a general awareness of heading. Precise aircraft heading (accurate to 1°) is shown digitally at the top of the display on the outer range ring. The range rings represent 25, 50, 75 and 100% of the range displayed in the upper right-hand corner.

10.2.1 Range display

The range of the display is shown in the upper right-hand corner in a slightly larger font than is found with the rest of the text on the display. The selectable display ranges are 20, 40, 60, and 80 nm.

10.2.2 Navigation information

Two data fields occupy the upper left-hand corner:

1. Current waypoint selected.
2. Bearing and range to current waypoint.

10.2.3 Time to Go (TTG)

The time to go in the mission is located at the bottom left of the SD screen. The TTG reflects how much time remains of the initial 10 minutes given for the mission.

10.2.4 Route of flight

Route of flight is drawn with waypoints as circles, the initial points as squares, and targets as triangles.

10.2.5 Location of FEBA

The FEBA is displayed as a blue line with semicircles evenly spaced along one side.

10.3 Attack Display (AD)

The AD is the host display for the A/G radar. The display shows radar ground returns as well as other important data. The A/G radar has two modes; real beam ground map (RBGM) mode for general ground mapping and SAR patch map for magnification of selected areas. This display also shows the current shootlist mode and lists each stage of the mission with their associated automation level.

10.4 Defensive Display (DD)

The DD is essentially a RWR indicator that displays munitions that are targeted at the ownship.

10.4.1 DD range and range rings

Range of the DD is shown in the upper-right hand corner of the display. Range rings represent that value (outer ring) and one-half the displayed range (inner ring). The range rings, like those used on the SD, have cardinal directions and digital heading displayed.

10.4.2 Countermeasure consumables

Chaff and flare counts are portrayed digitally at the bottom of the DD.

10.4.3 DD threat symbology

Detected munitions are indicated by "fans" extending from near the ownship symbol along the azimuth of the munitions. The location of the incoming missile is displayed with a red "M".

10.5 Headup display (HUD)

The HUD display format closely conforms to MIL-STD-1787 guidance with adjustments made for the unique SIRE environment and the addition of the shootlist. Figure 5 shows the likeness of the HUD that will be used for this experiment.

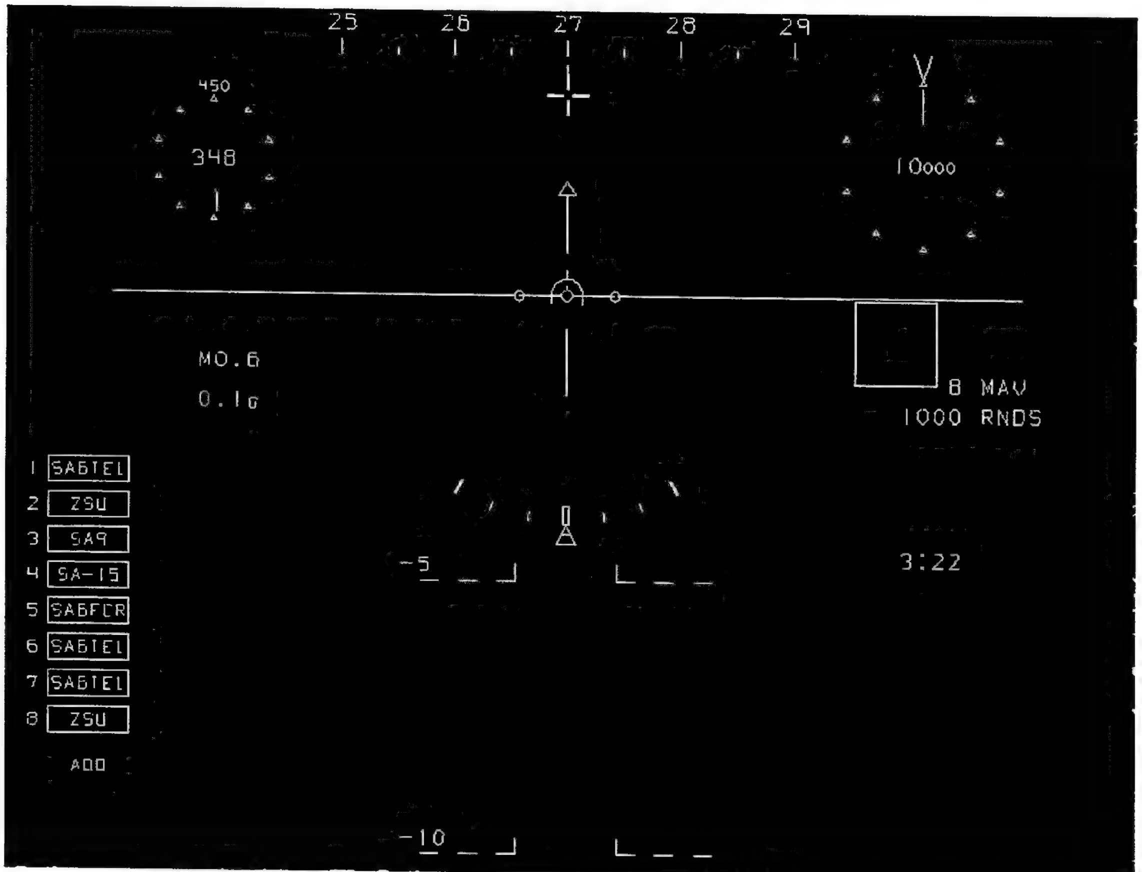


Figure 5. SIRE HUD display.

10.6 Switchology

Below are Figures of the switches on the stick and throttle with descriptions of their use in the tables that follow.

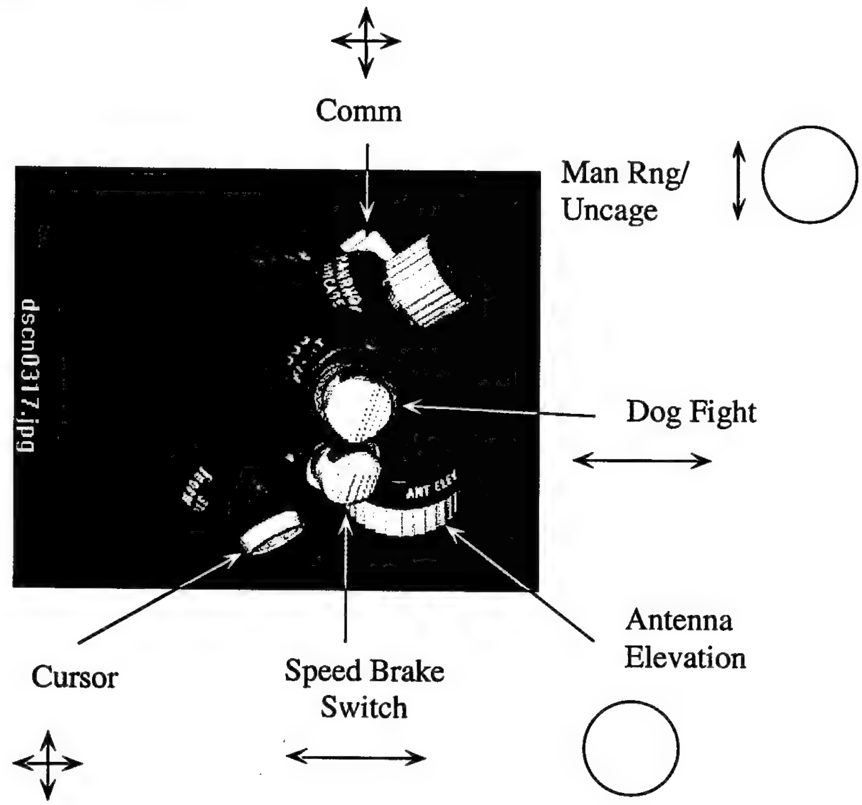


Figure 6. Throttle switches.

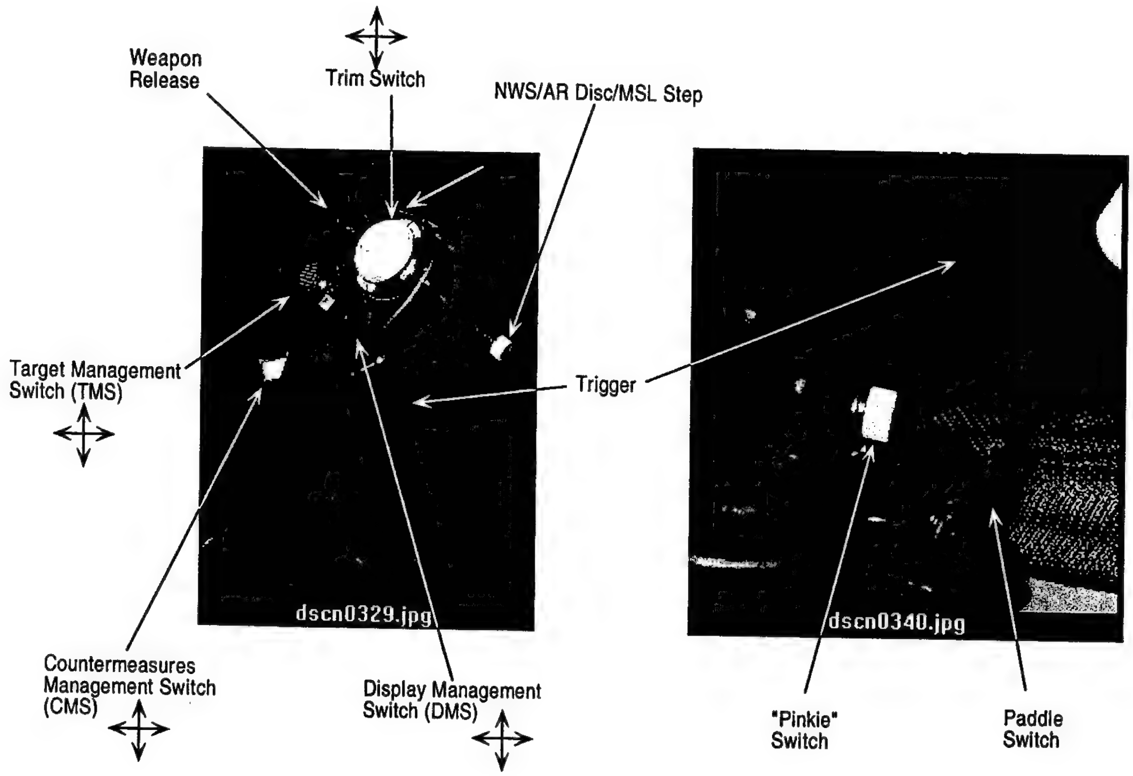


Figure 7. Stick switches.

10.6.1 Overall

Feature	Selections	Cockpit Switch	Feedback
Display in Command	AD or SD	DMS left/right	In command indication on AD or SD
Weapon Release		Weapon Release Button	Weapons released (auditory feedback)
Confirm		Missile Step	Auditory
Course Selection Toggle		Pinkie Switch	Course change on SD
Shootlist Scroll	One of eight slots	Speed Brake fore/aft	Scroll up/down on HUD
Shootlist mode	Varies by stage (see Table 3)	TMS left/right	Mode available on AD
Subjective measures item input	scales	Stick left/right	Highlighted pointer movement
Subjective measures /next item	Next/previous	Speed Brake fore/aft	Scroll to next/previous item

Table 6. Switchology for overall commands.

10.6.2 AD A/G Radar

Feature	Selections	Cockpit Switch	Feedback
Mode	RBGM& SAR	Uncage - Depress	Radar display changes format
Range	20/40/60/80	DMS – forward/increase, aft/decrease	20, 40, 60, or 80 on A/G AD, display scale changes
Cursor Movement	X-Y movement	Cursor	Cursor movement
SAR Map Size	2.3/1.7/.1	DMS – forward/increase, aft/decrease	2.3, 1.7 or .1 on A/G AD, display scale changes
Targeting	Designate/ Undesignate/ Ground Stabalized	TMS – Fore/Aft	Cursor status displayed on AD

Table 7. Switchology for AD A/G radar commands.

10.6.3 SD

Feature	Selections	Cockpit Switch	Feedback
Range	20/40/60/80	DMS – forward/increase, aft/decrease	20, 40, 60 or 80 on SD, display scale changes

Table 8. Switchology for SD commands.

10.6.4 DD

Feature	Selections	Cockpit Switch	Feedback
Range	5/10/20/40	CMS – Left/Right	Change in range indication and display
Countermeasures	Flares/Chaff	CMS – Fore/Aft	Auditory feedback

Table 9. Switchology for DD commands.

11 Subjective Measures

11.1 NASA TLX Definitions

NASA TASK LOAD INDEX (TLX) RATING SCALE DEFINITIONS

Dimension	Definition	Endpoints
Mental Demand	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?	*Low / High*
Physical Demand	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?	*Low / High*
Temporal Demand	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?	*Low / High*
Performance	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?	*Good / Poor*
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?	*Low / High*
Frustration	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?	*Low / High*

11.2 NASA TLX Rating Form

Automation Evaluation

Subject I.D.: ____________ Date: ______________ Session: ______________

NASA TLX (Task Load Index) Rating Scales

Instructions: For each of the scales presented below, please circle the hash-mark at the point which matches your experience with what you just completed. Please note that the "Performance" scale goes from "good" on the left to "poor" on the right.

Mental Demand

Low 0 25 50 75 100 High

Performance

Good 0 25 50 75 100 Poor

Physical Demand

Low 0 25 50 75 100 High

Effort

Low 0 25 50 75 100 High

Temporal Demand

Low 0 25 50 75 100 High

Frustration

Low 0 25 50 75 100 High

11.3 Situation Awareness Rating Form

Automation Evaluation I

SART (Situation Awareness Rating Technique)

Instructions: For each dimension below, please place a mark under the rating value that matches your experience with the taskt you just completed.

Dimensions	Rating						
	Low						High
	1	2	3	4	5	6	7
<u>Demands on Attentional Resources</u> Demands placed on your attentional resources by flying, navigation, and shooting tasks. How much of the task's instability, variability, and complexity affected your SA.							
<u>Supply of Attentional Resources</u> Think of your mental state while doing the task. Rating should reflect your degree of arousal, your spare mental capacity, your ability to concentrate, and your ability to divide attention across multiple tasks.							
<u>Understanding of the Situation</u> How your understanding and knowledge of the situation affects the task performance and SA. Please rate the quantity of information available to you, the quality of that information, and your familiarity with the task.							
<u>Overall SA</u> You should assume a broad perspective that takes into account your entire experience in the task, and to generate a single rating that you feel best represents your SA while performing the task.							

11.4 Trust and Confidence Rating Form

TRUST SCALE

Subject I.D.: __________ Date: ____________

Session: ____________ Block: ____________

How well can the systems behavior be predicted from moment to moment?

1	2	3	4	5	6	7	8	9	10

Can the system be counted on to do its job?

1	2	3	4	5	6	7	8	9	10

How much faith do you have that the system can cope with future events?

1	2	3	4	5	6	7	8	9	10

Overall, how much do you trust the system?

1	2	3	4	5	6	7	8	9	10

CONFIDENCE SCALE

Rate the confidence you had in your ability to complete the mission.

1	2	3	4	5	6	7	8	9	10

Rate the reliability of the automation, if present

1	2	3	4	5	6	7	8	9	10

11.5 Pre and Post Simulator Sickness Rating Form

SIRE Automation Evaluation

Subject I.D.: ________________ Date: __________________

PRE-EXPERIENCE COMFORT QUESTIONNAIRE

Instructions: Please circle the severity of any symptoms that apply to you right now.

	(0)	(1)	(2)	(3)
1. General Discomfort	None	Slight	Moderate	Severe
2. Fatigue	None	Slight	Moderate	Severe
3. Headache	None	Slight	Moderate	Severe
4. Eye Strain	None	Slight	Moderate	Severe
5. Difficulty Focusing	None	Slight	Moderate	Severe
6. Increased Salivation	None	Slight	Moderate	Severe
7. Sweating	None	Slight	Moderate	Severe
8. Nausea	None	Slight	Moderate	Severe
9. Difficulty Concentrating	None	Slight	Moderate	Severe
10. Fullness of Head	None	Slight	Moderate	Severe
11. Blurred Vision	None	Slight	Moderate	Severe
12. Dizzy (Eyes Open)	None	Slight	Moderate	Severe
13. Dizzy (Eyes Closed)	None	Slight	Moderate	Severe
14. Vertigo*	None	Slight	Moderate	Severe

*Vertigo refers to a loss of orientation with respect to upright (i.e., you don't know "which way is up")

15. Stomach Awareness**	None	Slight	Moderate	Severe

**Stomach awareness is usually used to indicate a feeling of discomfort which is just short of nausea

16. Burping	None	Slight	Moderate	Severe

Are there any other symptoms that you are experiencing right now? If so, please describe the symptom(s) and rate their severity below.

SIRE Automation Evaluation

Subject I.D.: ________________ Date: __________________

POST-EXPERIENCE COMFORT QUESTIONNAIRE

Instructions: Please circle the severity of any symptoms that apply to you right now; after experiencing the environment.

	(0)	(1)	(2)	(3)
1. General Discomfort	None	Slight	Moderate	Severe
2. Fatigue	None	Slight	Moderate	Severe
3. Headache	None	Slight	Moderate	Severe
4. Eye Strain	None	Slight	Moderate	Severe
5. Difficulty Focusing	None	Slight	Moderate	Severe
6. Increased Salivation	None	Slight	Moderate	Severe
7. Sweating	None	Slight	Moderate	Severe
8. Nausea	None	Slight	Moderate	Severe
9. Difficulty Concentrating	None	Slight	Moderate	Severe
10. Fullness of Head	None	Slight	Moderate	Severe
11. Blurred Vision	None	Slight	Moderate	Severe
12. Dizzy (Eyes Open)	None	Slight	Moderate	Severe
13. Dizzy (Eyes Closed)	None	Slight	Moderate	Severe
14. Vertigo*	None	Slight	Moderate	Severe

*Vertigo refers to a loss of orientation with respect to upright (i.e., you don't know "which way is up")

15. Stomach Awareness**	None	Slight	Moderate	Severe

**Stomach awareness is usually used to indicate a feeling of discomfort which is just short of nausea

16. Burping	None	Slight	Moderate	Severe

Are there any other symptoms that you are experiencing right now? If so, please describe the symptom(s) and rate their severity below.

APPENDIX B

INFORMATION PROTECTED BY THE PRIVACY ACT OF 1974

Informed Consent Document for Adaptive Interface Technology Development within the Synthesized Immersion Research Environment

1. Nature and Purpose: I have been asked to volunteer as a subject in the research study named above. The purpose is to assess performance within complex environments and to study human perception and performance with and/or without the use of multi-sensory (i.e., involving more than one sense, such as hearing and vision) and adaptive interfaces. Examples of this type of interface and the devices that may be used to generate this type of interface are described in the following paragraph. Testing is generally less than 2 hours. Occasionally, test periods that last 8 hours are employed, however, I will be given ample rest periods and, if desired, I may request additional rest. A total of 12 subjects will participate in this research study performed at AFRL/HECP, Wright Patterson AFB, Ohio 45433-7022.

2. Experimental Procedures: If I decide to participate, I may be asked to do any of the following: view a projected picture, target, or geometric design on a projection surface or video display; view a simulated aircraft; view a static or moving display or helmet-mounted display system, listen to a simulated target, aircraft, or other nonspeech or speech audio signals, track targets using hand and/or head movements; control a simulated aircraft using hand movements; complete questionnaires regarding symptoms of fatigue, subjective levels of workload, symptoms of simulator sickness. I may have skin electrodes attached to my head or neck during testing as well as respiration monitoring devices. In addition, I may be asked to control a flight simulator using multi-sensory (i.e., involving more than one sense, such as hearing and vision) and partially immersive (i.e., virtual reality) control and display devices. Testing may be done in normal lighting or low lighting conditions. Normally, I will be seated during testing and will respond verbally or with a response button or stick. I will be asked to perform postural stability tests (such as heel-to-toe walking and standing on one leg), before and after the experiment, to determine the extent to which I have experienced symptoms associated with simulator sickness.

3. Discomfort and Risks: Viewing visual representations of motion has been found, in some cases, to induce discomfort, nausea, dizziness, and headaches, commonly referred to as motion sickness. Although the displays and the experiment have been designed to minimize these risks, I am aware that subjects participating in similar experiments have noted these side effects. I will let the experimenter know if I am experiencing these effects. If I, or the medical monitor, determine that I cannot transport myself back to my residence at the termination of this experiment, I will be offered one-way transportation back to my residence on the day of the experiment. Some experiments may involve flickering lights. Flickering lights (for example, televisions and strobe lights) have been known to cause seizure activity in a small percentage of the population with seizure related disorders. I have disclosed to the investigator if I have a seizure-related disorder or a history of seizure activity of any kind.

4. Precautions for Female Subjects: There are no special precautions for female subjects.

5. Benefits:
I will not receive any known medical benefits resulting from participation in this experiment. However, my participation in this study will provide me with an opportunity to experience state-of-the-art virtual reality technology. And, I am encouraged to provide the experimenter with feedback about the experiment so that my concerns can be considered in future investigations.

6. Entitlements and Confidentiality: Records of my participation in this study may only be disclosed according to federal law, including the Federal Privacy Act, 5 U.S.C. 552a, and its implementing regulations.

I understand my entitlements to medical and dental care and/or compensation in the event of injury are governed by federal laws and regulations, and that if I desire further information I may contact the base legal office (88ABW/JA - phone 257-6142).

If an unanticipated event (medical misadventure) occurs during my participation in this study, I will be informed. If I am not competent at the time to understand the nature of the event, such information will be brought to the attention of my next of kin.

The decision to participate in this research is completely voluntary on my part. No one has coerced or intimidated me into participating in this program. I am participating because I want to. Scott Galster, or his representative, ________________, has adequately answered any and all questions I have about this study, my participation, and the procedures involved. I understand that Scott Galster, or his representative, ______________, will be available to answer any questions concerning procedures throughout this study. I understand that if significant new findings develop during the course of this research, which may relate to my decision to continue participation, I will be informed. I further understand that I may withdraw this consent at any time and discontinue further participation in this study without prejudice to my entitlements. I also understand that the medical monitor of this study may terminate my participation in this study if she or he feels this to be in my best interest.

____ ____________
VOLUNTEER SIGNATURE AND SSN (optional) DATE

______________________________ ____________
INVESTIGATOR SIGNATURE DATE

______________________________ ____________
WITNESS SIGNATURE DATE

Privacy Act Statement

Authority: We are requesting disclosure of personal information, to include your Social Security Number. Researchers are authorized to collect personal information (including social security numbers) on research subjects under The Privacy Act-5 USC 552a, 10 USC 55, 10 USC 8013, 32 CFR 219, 45 CFR Part 46, and EO 9397, November 1943 (SSN).

Purpose: It is possible that latent risks or injuries inherent in this experiment will not be discovered until some time in the future. The purpose of collecting this information is to aid researchers in locating you at a future date if further disclosures are appropriate.

Routine Uses: Information (including name and SSN) may be furnished to Federal, State and local agencies for any uses published by the Air Force in the Federal Register, 52 FR 16431, to include, furtherance of the research involved with this study and to provide medical care.

Disclosure: Disclosure of the requested information is voluntary. No adverse action whatsoever will be taken against you, and no privilege will be denied you based on the fact you do not disclose this information. However, your participation in this study may be impacted by a refusal to provide this information.

COGNITIVE SCIENCE LABORATORY
250 O'Boyle Hall
The Catholic University of America
Washington DC 20064

Tel: (202) 319-5825
Fax: (202) 319-4456

CONSENT FORM

I state that I am over eighteen (18) years of age and wish to participate in the study entitled *An Examination of Complex Human-Machine System Performance under Multiple Levels and Stages of Automation.* This work is being carried out jointly by the Air Force Research Laboratory, Human Effectiveness Directorate, Crew Systems Interface Division, Human Interface Systems Branch, Wright-Patterson Air Force Base, and The Catholic University of America, Washington, DC.

Research Objectives

The purpose of this research is to examine human-machine performance differences under several automated conditions during a tactical mission. This research will be conducted using the Synthesized Immersion Research Environment (SIRE) located at AFRL/HECP, Wright-Patterson AFB, OH. By participating, I will contribute to research examining the nature of human-interaction with automation. Data from these studies will be used to partially fulfill the requirements for a doctoral dissertation that will be submitted to The Catholic University of America by the experimenter.

Procedures

The following will occur during my participation: I will be briefed and trained by a Subject Matter Expert on how to operate the flight simulator and instructed on the use of each avionics system. I will then complete thirty two (32) combat missions, each lasting ten (10) minutes, over a two-day period. I will be asked to complete brief questionnaires after each mission. Rest periods will be normally be given after eight missions but will be granted at any time upon request.

I understand that this study has a small risk of inducing motion sickness as evidenced by participants in similar studies. If at any point I experience discomfort, I may request that testing be stopped. I also understand that I am free to end my participation in this study at any time, without penalty.

Compensation

This study is voluntary and I will not receive monetary compensation.

Your right to confidentiality

If I have any questions concerning participation in this experiment, I will address them to the experimenter at this time or at any time during the experiment. I understand that I am free to deny answers to specific items or questions in the questionnaires.

The data obtained from the participants will be coded before transcription to preserve confidentiality. Only the principal investigator will have access to the codes, which will be kept in a secure manner. Participants will not be identified by name in any report of the results.

Records of my participation in this study may only be disclosed according to federal law, including the Federal Privacy Act, 5 U.S.C. 552a, and its implementing regulations.

If I have any concerns at the end of this study related to the procedures involved or how the study was conducted I should direct them to the experimenter at the conclusion of the experiment. I may also contact the Secretary of the Committee for the Protection of Human Subjects at (202) 319-5218. The Secretary is located in the Office of Sponsored Research on the CUA campus.

I agree to participate in this research study.

I understand and agree to the terms outlined in this document and have been provided a copy if requested.

Signature of Participant ________________________________ Date________

Signature of Experimenter ______________________________ Date________

Subject #____________
Date____________

PILOT BACKGROUND FORM: AUTOMATION STUDY

Name: ____

Organization: ___________________________________

Address: ______________________________________

____

Phone: _______________________________________

Email: _______________________________________

Age (yrs): ________ **DOB** __________ **Gender:** **Male** ☐ **Female** ☐

Handedness **Right** ☐ **Left** ☐ **Right/Left** ☐

Vision/Hearing **Normal** ☐ **Corrected to Normal** ☐ **Deficient** ☐

If Deficient, please describe: ______________________________

____

Total Flying Time: ________ **Combat Flying Time:** ________

Total Jet Time: ________ **Military Flying Time:** ________

Current Aircraft and Hours:	**Other Aircraft and Hours:**

Any other relevant background information, e.g. experience with other simulation experiments?

APPENDIX C

Training Checklist

Introduction

- Introduce HUD Symbology
- Maintenance of Flight Parameters
- Low level Flight

Static Training

- **Switches and Displays**
  - Switches
    - Speed Brake
    - TMS
    - DMS
    - CMS
  - Displays
    - *Defensive Display*
      - Range
      - Chaff
      - Flares
    - *Situation Display*
      - Range
      - FEBA
      - Triangles/Groups
      - Waypoints
      - TTG
    - *Attack Display*
      - Shootlist Modes
      - Automation Condition
      - Cursor Designation
      - Range to Target
    - *HUD*
      - CDM
      - Airspeed
      - Altitude
      - Pitch
      - Heading
      - CDI
      - Weapons Count
      - Timer

Range
Snow Plow/Ground Stabalized/Designated
Cursor Movement
Groups
Making a SAR Patch Map
- Ground Stabalized
- Designated

Cursor Movement in SAR
Movement of SAR map on Situation Display

Stage Tasks

<u>Stage I</u>

- Shoot List (show all)
  - Add
  - Remove
  - Replace
  - Insert
  - Move
  - Deploy
  - Building a Shootlist
  - Confirm Button

<u>Stage II</u>

- Move (Cut & Paste)
- Keeping Groups Seperated
- Confirm Button

<u>Stage III</u>

- Random Course Inserted
- Target Locations (Deploy/Speed Brake)
- Course Toggle (pinkie switch)
- Best Course for Shootlist
- Confirm Button

<u>Stage IV</u>

- Designating Target
- TD Box
- Target Locater Line/Degrees off Nose
- Missile Step
- Undesignate/Toggle/Designate
- Circle on Shootlist

Subjective Measures

- Stick/Speed Brake inputs
- TLX Definitions
- TLX Scales (point out performance scale)
- SART
- Trust & Confidence
- Global Score
- Confirm Button

Dynamic Training

- Manual
  - Walk-through
  - Flight Parameters
  - Reminders
  - Missile Lock Tone

Automation Training

- Point out differences
- Stage I
- Stage II
  - Wait for Automation
- Stage III
  - Wait for Automation
- Stage IV
  - Launch Codes
  - Switch Differences

Training Objectives

Training Exercise	Flight Control	Targets per Group	Stage				Training Objective
			I	II	III	IV	
Cockpit Dynamics	Static	0	X	X	X	X	Let pilot "fly" around to get acquainted with the simulator capabilities, dynamics and responses to flight control inputs. Encourage low-level evasive maneuvering.
Task/Switches	Static	4	M	M	M	M	Introduction to task objectives for each stage. Pilot is not flying or otherwise controlling aircraft. Pilot is instructed on switchology for each stage. Pilot is also trained on modes and functions of the shootlist, HUD symbology, display symbology, functions and capabilities of each display. This session can be repeated until the pilot indicates they are ready to advance to subsequent training trials. The pilot is instructed to make errors and taught how to correct for the errors. The pilot is instructed on inputs for Subjective ratings.

Training Exercise	Flight Control	Targets per Group	Stage				Training Objective
			I	II	III	IV	
1	Dynamic	5	M	M	M	M	Pilot performs the tasks manually in all stages while flying and maintaining flight parameters. In this and all training sessions an experimenter will "walk" the pilot through the training session pointing out errors (if any).
2	Dynamic	5	M	M	M	M	Second trial with all manual. Objectives repeated.

Training Exercise	Flight Control	Targets per Group	Stage				Training Objective
			I	II	III	IV	
3	Dynamic	5	A	M	M	M	Introduction of Automation in Stage I - Identification of all targets. Surface to Air threats are free to fire in this and next two training session while pilot is in the gaming area (Stage IV); allows the pilot to hear the tone for an incoming missile and practice avoidance maneuvers.
4	Dynamic	6	M	A	M	M	Introduction of Automation in Stage II - Shootlist prioritization. Pilot instructed to wait until the automation is finished sorting list before initiating any changes (if any).
5	Dynamic	6	M	M	A	M	Introduction of Automation in Stage III - Pilot is instructed to wait until the automation has made it's recommended flight path before initiating any changes (if any)

Training Exercise	**Flight Control**	**Targets per Group**	**Stage**				**Training Objective**
			I	**II**	**III**	**IV**	
6	Dynamic	7	M	M	M	A	Introduction of Automation in Stage IV - Pilot is shown shoot cues for missile launch and instructed on the automatic de-selection and designation of the next target in the shootlist.
7	Dynamic	8	A	A	A	A	Automation at all Stages
8	Dynamic	8	A	A	A	A	Automation at all Stages
Notes							Pilot is free to run through the satatic training again. Pilot is required to do static training at the beginning of subsequent sessions. Pilot can opt to do more than one static training session, data collection begins/resumes when pilot indicates they are ready.

APPENDIX D

Automation Study 1
Post-Experimental Questionnaire

Subject:__________ Date:__________

The experiment you just completed compared manual and automated aided performance at different stages. Please answer the questions below regarding your experiences.

1. On a scale of 1-10, please indicate your agreement with the statement "Overall, the automation was helpful, when present, in completing the mission". (1= disagree greatly, 5= neither disagree nor agree, 10= agree greatly.

______

2. Please rank, in order, the stage that the automation was most helpful.

	Stage
Most Helpful	
Least Helpful	

3. In regards to completing the mission successfully, please indicate the best configuration for each stage.

Stage	Automated or Manual
I	
II	
III	
IV	

4. On a scale of 1-10, please indicate how much the automation improved your Situation Awareness. (1= greatly decreased my SA, 5= did not affect my SA, 10= greatly improved my SA.

Stage	Situation Awareness
I	
II	
III	
IV	

5. On a scale of 1-10, please indicate your reliance on the automation. (1= did not rely at all on the automation, 5= neutral, 10= relied greatly on the automation

Stage	Reliance
I	
II	
III	
IV	

6. On a scale of 1-10, please indicate the impact on your performance the automation had at each stage if it was unreliable (1= unreliability did not impact performance at all, 5= neutral, 10= unreliability impacted performance greatly.

Stage	Impact on Performance
I	
II	
III	
IV	

7. If only two stages could be automated, which stages would you choose? Why?

8. Please add any comments you may have regarding this experiment.

APPENDIX E

ANOVA Tables for Stage I Measures

ANOVA table for Stage I Primary Task Scores

Source	df	MS	F	p
Stage I (A)	1	157262.00	104.14	<.05
Workload (B)	1	32738.40	19.26	<.05
A × B	1	32738.40	17.95	<.05
S × A	7	1510.03		
S × B	7	1699.54		
S × A × B	7	1824.09		

ANOVA table for Stage I Secondary Task Scores

Source	df	MS	F	p
Stage I (A)	1	2727.66	2.44	ns
Workload (B)	1	2549.55	2.71	ns
A × B	1	1039.35	12.11	<.05
S × A	7	1118.39		
S × B	7	942.31		
S × A × B	7	85.83		

ANOVA table for Stage I Altitude Deviations

Source	df	MS	F	p
Stage I (A)	1	132331.00	0.71	ns
Workload (B)	1	34384.00	0.52	ns
A × B	1	19994.00	0.35	ns
S × A	7	186596.00		
S × B	7	66103.50		
S × A × B	7	57949.90		

ANOVA table for Stage I Airspeed Deviations

Source	df	MS	F	p
Stage I (A)	1	43.61	0.39	ns
Workload (B)	1	130.05	2.63	ns
A × B	1	9.11	0.12	ns
S × A	7	112.43		
S × B	7	49.47		
S × A × B	7	73.78		

ANOVA table for Stage I Cross Track Errors

Source	df	MS	F	p
Stage I (A)	1	0.01	0.23	ns
Workload (B)	1	0.04	1.60	ns
A × B	1	0.01	0.57	ns
S × A	7	0.06		
S × B	7	0.02		
S × A × B	7	0.02		

ANOVA table for Stage I Track Angle Errors

Source	df	MS	F	p
Stage I (A)	1	13.50	5.14	ns
Workload (B)	1	11.95	2.89	ns
A × B	1	5.38	7.20	<.05
S × A	7	2.63		
S × B	7	4.14		
S × A × B	7	0.75		

ANOVA table for Stage I Total Scores

Source	df	MS	F	p
Stage I (A)	1	118567.00	31.19	<.05
Workload (B)	1	17015.70	3.72	ns
A × B	1	4544.20	20.65	<.05
S × A	7	3800.89		
S × B	7	4568.62		
S × A × B	7	2200.48		

ANOVA table for Stage I Confirm Button Press Time

Source	df	MS	F	p
Stage I (A)	1	52972.50	32.55	<.05
Workload (B)	1	614.73	2.86	ns
A × B	1	2837.05	21.54	<.05
S × A	7	1627.37		
S × B	7	214.82		
S × A × B	7	131.69		

ANOVA table for Stage I Last Switch

Source	df	MS	F	p
Stage I (A)	1	81607.30	63.55	<.05
Workload (B)	1	1496.73	6.52	<.05
A × B	1	2740.33	24.34	<.05
S × A	7	1284.11		
S × B	7	229.68		
S × A × B	7	112.60		

ANOVA table for Stage I Elapsed Time

Source	df	MS	F	p
Stage I (A)	1	14.38	4.25	ns
Workload (B)	1	1.05	0.12	ns
A × B	1	0.06	0.02	ns
S × A	7	3.39		
S × B	7	8.97		
S × A × B	7	3.27		

ANOVA Tables for Stage II Measures

ANOVA table for Stage II Primary Task Scores

Source	df	MS	F	p
Stage I (A)	1	6006.25	7.72	<.05
Stage II (B)	1	225.00	0.72	ns
Workload (C)	1	3306.25	29.16	<.05
A × B	1	56.25	0.11	ns
A × C	1	1225.00	7.00	<.05
B × C	1	306.25	0.77	ns
A × B × C	1	100.00	0.14	ns
S × A	7	777.68		
S × B	7	310.71		
S × C	7	113.39		
S × A × B	7	527.68		
S × A × C	7	175.00		
S × B × C	7	399.11		
S × A × B × C	7	707.14		

ANOVA table for Stage II Secondary Task Scores

Source	df	MS	F	p
Stage I (A)	1	9420.56	4.02	ns5
Stage II (B)	1	187.98	0.14	ns
Workload (C)	1	0.05	0.00	ns
A × B	1	1020.87	0.65	ns
A × C	1	688.57	1.21	ns
B × C	1	1155.08	1.00	ns
A × B × C	1	210.23	0.44	ns
S × A	7			
S × B	7			
S × C	7			
S × A × B	7			
S × A × C	7			
S × B × C	7			
S × A × B × C	7			

ANOVA table for Stage II Altitude Deviations

Source	df	MS	F	p
Stage I (A)	1	342767.00	7.19	<.05
Stage II (B)	1	324891.00	1.21	ns
Workload (C)	1	688.53	0.03	ns
A × B	1	144527.00	1.16	ns
A × C	1	6290.20	0.08	ns
B × C	1	40061.40	0.31	ns
A × B × C	1	11633.20	0.78	ns
S × A	7	47677.00		
S × B	7	268281.00		
S × C	7	20685.00		
S × A × B	7	124855.00		
S × A × C	7	74073.50		
S × B × C	7	128963.00		
S × A × B × C	7	14997.00		

ANOVA table for Stage II Airspeed Deviations

Source	df	MS	F	p
Stage I (A)	1	2213.33	9.75	<.05
Stage II (B)	1	691.25	2.62	ns
Workload (C)	1	105.86	0.74	ns
A × B	1	0.54	0.00	ns
A × C	1	307.89	1.51	ns
B × C	1	99.66	0.31	ns
A × B × C	1	13.53	0.08	ns
S × A	7	226.91		
S × B	7	263.82		
S × C	7	142.78		
S × A × B	7	161.93		
S × A × C	7	204.02		
S × B × C	7	321.93		
S × A × B × C	7	162.59		

ANOVA table for Stage II Cross Track Errors

Source	df	MS	F	p
Stage I (A)	1	0.60	3.69	ns
Stage II (B)	1	0.06	0.68	ns
Workload (C)	1	0.00	0.01	ns
A × B	1	0.39	3.09	ns
A × C	1	0.03	0.47	ns
B × C	1	0.58	9.02	<.05
A × B × C	1	0.01	0.09	ns
S × A	7	0.16		
S × B	7	0.09		
S × C	7	0.14		
S × A × B	7	0.13		
S × A × C	7	0.07		
S × B × C	7	0.06		
S × A × B × C	7	0.13		

ANOVA table for Stage II Track Angle Errors

Source	df	MS	F	p
Stage I (A)	1	16.31	1.64	ns
Stage II (B)	1	8.51	5.87	<.05
Workload (C)	1	0.42	0.06	ns
A × B	1	23.23	4.50	ns
A × C	1	1.59	0.37	ns
B × C	1	12.38	2.56	ns
A × B × C	1	3.46	0.95	ns
S × A	7	9.96		
S × B	7	1.45		
S × C	7	6.90		
S × A × B	7	5.17		
S × A × C	7	4.28		
S × B × C	7	4.83		
S × A × B × C	7	3.63		

ANOVA table for Stage II Total Scores

Source	df	MS	F	p
Stage I (A)	1	30471.00	7.94	<.05
Stage II (B)	1	824.31	0.54	ns
Workload (C)	1	3281.13	4.56	ns
A × B	1	597.85	0.18	ns
A × C	1	76.73	0.10	ns
B × C	1	2650.86	1.86	ns
A × B × C	1	600.22	0.94	ns
S × A	7	3839.25		
S × B	7	1537.79		
S × C	7	719.32		
S × A × B	7	3416.37		
S × A × C	7	762.28		
S × B × C	7	1422.12		
S × A × B × C	7	640.30		

ANOVA table for Stage II Confirm Button Press Time

Source	df	MS	F	p
Stage I (A)	1	419.43	1.59	ns
Stage II (B)	1	1.98	0.01	ns
Workload (C)	1	1472.74	10.37	<.05
A × B	1	141.41	0.74	ns
A × C	1	60.84	0.28	ns
B × C	1	5.37	0.02	ns
A × B × C	1	70.82	0.45	ns
S × A	7	263.06		
S × B	7	262.12		
S × C	7	141.60		
S × A × B	7	190.44		
S × A × C	7	216.87		
S × B × C	7	227.14		
S × A × B × C	7	156.52		

ANOVA table for Stage II Last Switch

Source	df	MS	F	p
Stage I (A)	1	1283.53	9.15	<.05
Stage II (B)	1	39.11	0.08	ns
Workload (C)	1	1161.41	5.76	<.05
A × B	1	1023.06	2.14	ns
A × C	1	0.93	0.00	ns
B × C	1	6.11	0.02	ns
A × B × C	1	156.01	2.14	ns
S × A	7	140.35		
S × B	7	491.57		
S × C	7	201.61		
S × A × B	7	479.18		
S × A × C	7	393.34		
S × B × C	7	338.84		
S × A × B × C	7	73.00		

ANOVA table for Stage II Elapsed Time

Source	df	MS	F	p
Stage I (A)	1	0.06	0.12	ns
Stage II (B)	1	1.60	1.14	ns
Workload (C)	1	4.53	2.00	ns
A × B	1	3.71	1.25	ns
A × C	1	2.24	1.34	ns
B × C	1	0.29	0.10	ns
A × B × C	1	9.63	3.18	ns
S × A	7	1.35		
S × B	7	1.40		
S × C	7	2.27		
S × A × B	7	2.96		
S × A × C	7	1.67		
S × B × C	7	2.87		
S × A × B × C	7	3.03		

ANOVA Tables for Stage III Measures

ANOVA table for Stage III Primary Task Scores

Source	df	MS	F	p
Stage II (A)	1	351.56	0.88	ns
Stage III (B)	1	0.00	0.00	ns
Workload (C)	1	1406.25	9.33	<.05
A × B	1	791.02	2.33	ns
A × C	1	87.89	0.26	ns
B × C	1	87.89	0.64	ns
A × B × C	1	87.89	0.20	ns
S × A	7	401.79		
S × B	7	251.12		
S × C	7	150.67		
S × A × B	7	339.01		
S × A × C	7	339.01		
S × B × C	7	138.11		
S × A × B × C	7	439.45		

ANOVA table for Stage III Secondary Task Scores

Source	df	MS	F	p
Stage II (A)	1	2021.56	1.49	ns
Stage III (B)	1	259.04	0.07	ns
Workload (C)	1	475.63	0.30	ns
A × B	1	0.21	0.00	ns
A × C	1	8591.44	5.87	<.05
B × C	1	1.13	0.00	ns
A × B × C	1	3158.85	6.78	<.05
S × A	7	1356.72		
S × B	7	3757.17		
S × C	7	1563.60		
S × A × B	7	1463.51		
S × A × C	7	1464.78		
S × B × C	7	3262.63		
S × A × B × C	7	466.13		

ANOVA table for Stage III Altitude Deviations

Source	df	MS	F	p
Stage II (A)	1	24805.70	0.10	ns
Stage III (B)	1	81008.50	0.27	ns
Workload (C)	1	797.70	0.00	ns
A × B	1	16792.90	0.08	ns
A × C	1	167462.00	0.68	ns
B × C	1	540.36	0.00	ns
A × B × C	1	443711.00	10.08	<.05
S × A	7	240449.00		
S × B	7	297708.00		
S × C	7	422987.00		
S × A × B	7	206608.00		
S × A × C	7	248067.00		
S × B × C	7	494768.00		
S × A × B × C	7	44040.30		

ANOVA table for Stage III Airspeed Deviations

Source	df	MS	F	p
Stage II (A)	1	246.29	2.14	ns
Stage III (B)	1	181.00	0.97	ns
Workload (C)	1	2.80	0.04	ns
A × B	1	93.25	2.90	ns
A × C	1	59.58	0.98	ns
B × C	1	0.18	0.00	ns
A × B × C	1	131.60	0.39	ns
S × A	7	114.89		
S × B	7	186.36		
S × C	7	73.00		
S × A × B	7	32.11		
S × A × C	7	61.09		
S × B × C	7	43.43		
S × A × B × C	7	339.48		

ANOVA table for Stage III Cross Track Errors

Source	df	MS	F	p
Stage II (A)	1	0.04	0.57	ns
Stage III (B)	1	0.00	0.01	ns
Workload (C)	1	0.00	0.05	ns
A × B	1	0.14	3.19	ns
A × C	1	0.13	1.45	ns
B × C	1	0.00	0.08	ns
A × B × C	1	0.11	1.98	ns
S × A	7	0.07		
S × B	7	0.13		
S × C	7	0.03		
S × A × B	7	0.04		
S × A × C	7	0.09		
S × B × C	7	0.04		
S × A × B × C	7	0.05		

ANOVA table for Stage III Track Angle Errors

Source	df	MS	F	p
Stage II (A)	1	7.97	3.41	ns
Stage III (B)	1	0.22	0.01	ns
Workload (C)	1	5.69	1.53	ns
A × B	1	0.00	0.00	ns
A × C	1	44.13	7.67	<.05
B × C	1	0.05	0.00	ns
A × B × C	1	7.29	5.08	ns
S × A	7	2.34		
S × B	7	27.46		
S × C	7	3.73		
S × A × B	7	12.10		
S × A × C	7	5.78		
S × B × C	7	11.43		
S × A × B × C	7	1.43		

ANOVA table for Stage III Total Scores

Source	df	MS	F	p
Stage II (A)	1	4059.18	1.82	ns
Stage III (B)	1	259.04	0.08	ns
Workload (C)	1	3517.54	1.76	ns
A × B	1	765.33	0.30	ns
A × C	1	10417.30	5.78	<.05
B × C	1	108.97	0.03	ns
A × B × C	1	2192.92	36.16	<.05
S × A	7	2233.58		
S × B	7	3309.29		
S × C	7	1993.26		
S × A × B	7	2540.55		
S × A × C	7	1802.47		
S × B × C	7	3821.03		
S × A × B × C	7	60.65		

ANOVA table for Stage III Confirm Button Press Time

Source	df	MS	F	p
Stage II (A)	1	35.62	0.57	ns
Stage III (B)	1	20.29	1.39	ns
Workload (C)	1	5.62	0.14	ns
A × B	1	22.67	0.60	ns
A × C	1	9.78	0.33	ns
B × C	1	0.31	0.01	ns
A × B × C	1	52.23	1.15	ns
S × A	7	62.33		
S × B	7	14.56		
S × C	7	40.36		
S × A × B	7	37.64		
S × A × C	7	30.03		
S × B × C	7	47.77		
S × A × B × C	7	45.54		

ANOVA table for Stage III Last Switch

Source	df	MS	F	p
Stage II (A)	1	18.95	0.27	ns
Stage III (B)	1	383.60	4.71	ns
Workload (C)	1	120.18	12.80	<.05
A × B	1	223.76	12.63	<.05
A × C	1	87.52	1.68	ns
B × C	1	3.37	0.19	ns
A × B × C	1	0.13	0.00	ns
S × A	7	70.47		
S × B	7	81.36		
S × C	7	9.39		
S × A × B	7	17.72		
S × A × C	7	52.25		
S × B × C	7	17.89		
S × A × B × C	7	48.51		

ANOVA table for Stage III Elapsed Time

Source	df	MS	F	p
Stage II (A)	1	0.90	1.21	ns
Stage III (B)	1	0.86	0.45	ns
Workload (C)	1	0.25	1.37	ns
A × B	1	0.26	0.37	ns
A × C	1	0.20	0.49	ns
B × C	1	0.00	0.00	ns
A × B × C	1	0.00	0.00	ns
S × A	7	0.74		
S × B	7	1.89		
S × C	7	0.18		
S × A × B	7	0.73		
S × A × C	7	0.41		
S × B × C	7	1.13		
S × A × B × C	7	0.53		

ANOVA Tables for Stage IV Measures

ANOVA table for Stage IV Primary Task Scores

Source	df	MS	F	p
Stage III (A)	1	76.56	0.14	ns
Stage IV (B)	1	826.56	0.64	ns
Workload (C)	1	1701.56	3.21	ns
A × B	1	1501.56	7.45	<.05
A × C	1	126.56	0.21	ns
B × C	1	351.56	1.09	ns
A × B × C	1	39.06	0.05	ns
S × A	7	533.71		
S × B	7	1297.99		
S × C	7	530.13		
S × A × B	7	201.56		
S × A × C	7	612.28		
S × B × C	7	322.99		
S × A × B × C	7	839.06		

ANOVA table for Stage IV Average Range

Source	df	MS	F	p
Stage III (A)	1	458.48	19.09	<.05
Stage IV (B)	1	862.62	19.06	<.05
Workload (C)	1	3.89	0.08	ns
A × B	1	142.99	8.34	<.05
A × C	1	57.48	2.07	ns
B × C	1	12.56	0.79	ns
A × B × C	1	67.33	3.27	ns
S × A	7	23.78		
S × B	7	45.38		
S × C	7	46.16		
S × A × B	7	16.79		
S × A × C	7	27.53		
S × B × C	7	15.45		
S × A × B × C	7	20.32		

ANOVA table for Stage IV Group 1 Range

Source	df	MS	F	p
Stage III (A)	1	859.60	24.51	<.05
Stage IV (B)	1	350.14	4.57	ns
Workload (C)	1	113.88	1.01	ns
A × B	1	219.83	8.95	<.05
A × C	1	38.90	0.67	ns
B × C	1	0.16	0.01	ns
A × B × C	1	30.74	1.19	ns
S × A	7	34.75		
S × B	7	77.00		
S × C	7	113.76		
S × A × B	7	24.05		
S × A × C	7	58.42		
S × B × C	7	24.43		
S × A × B × C	7	25.42		

ANOVA table for Stage IV Group 2 Range

Source	df	MS	F	p
Stage III (A)	1	182.40	2.94	ns
Stage IV (B)	1	1602.30	44.20	<.05
Workload (C)	1	45.22	0.76	ns
A × B	1	82.62	1.25	ns
A × C	1	79.68	1.03	ns
B × C	1	56.11	0.93	ns
A × B × C	1	118.08	4.43	ns
S × A	7	61.62		
S × B	7	35.36		
S × C	7	58.89		
S × A × B	7	65.97		
S × A × C	7	76.86		
S × B × C	7	59.67		
S × A × B × C	7	25.65		

ANOVA Tables for Global Measures

ANOVA table for Global Primary Task Scores

Source	df	MS	F	p
Stage I (A)	1	279907.00	38.47	<.05
Workload (B)	1	93215.70	19.90	<.05
A × B	1	82836.00	12.47	<.05
S × A	7	7276.55		
S × B	7	4684.25		
S × A × B	7	6642.51		

ANOVA table for Global Secondary Task Scores

Source	df	MS	F	p
Stage I (A)	1	8510.06	0.91	ns
Workload (B)	1	877.64	0.25	ns
A × B	1	126.56	0.12	ns
S × A	7	9312.23		
S × B	7	3543.45		
S × A × B	7	1075.91		

ANOVA table Global Scores

Source	df	MS	F	p
Stage I (A)	1	386029.00	14.88	<.05
Workload (B)	1	76003.60	5.17	ns
A × B	1	76486.80	8.68	<.05
S × A	7	25934.50		
S × B	7	14693.40		
S × A × B	7	8808.58		

ANOVA table for Average NASA-TLX Ratings

Source	df	MS	F	p
Stage I (A)	1	12705.10	22.23	<.05
Workload (B)	1	607.81	13.25	<.05
A × B	1	747.84	5.38	ns
S × A	7	571.57		
S × B	7	45.88		
S × A × B	7	139.05		

ANOVA table for Overall Situation Awareness Ratings

Source	df	MS	F	p
Stage I (A)	1	49.00	5.70	<.05
Workload (B)	1	6.89	4.25	ns
A × B	1	9.77	5.83	<.05
S × A	7	8.59		
S × B	7	1.62		
S × A × B	7	1.68		

ANOVA table for Overall Trust Ratings

Source	df	MS	F	p
Stage I (A)	1	5235.83	5.90	<.05
Workload (B)	1	291.37	0.92	ns
A × B	1	308.44	2.33	ns
S × A	7	886.47		
S × B	7	318.40		
S × A × B	7	132.29		

ANOVA table for Confidence Ratings

Source	df	MS	F	p
Stage I (A)	1	7896.13	8.62	<.05
Workload (B)	1	186.92	1.53	ns
A × B	1	636.38	1.64	ns
S × A	7	916.21		
S × B	7	121.85		
S × A × B	7	387.36		

ANOVA table for Reliability Ratings

Source	df	MS	F	p
Stage I (A)	1	4226.47	8.20	<.05
Workload (B)	1	433.75	1.90	ns
A × B	1	225.65	3.32	ns
S × A	7	515.60		
S × B	7	228.77		
S × A × B	7	67.98		

BIBLIOGRAPHY

Baddeley, A.D., (1996). Working memory. Oxford, U.K.

Bailey, L.L., & Thompson, R.C. (2001). The TLX: One or more constructs. In Proceedings of the 11th International Symposium of Aviation Psychology (1-4). Columbus, OH: Ohio State University.

Bainbridge, L. (1983). Ironies of automation. Automatica,19, 775-779.

Billings, C. E. (1991). Human-centered aircraft automation philosophy: A concept and guidelines (Technical Memo No. 103885). Moffett Field, CA: NASA Ames Research Center.

Billings, C. E. (1997). Aviation Automation: The Search for a Human-Centered Approach. Mahwah, NJ: Erlbaum.

Billings, C.E., Lauber, J.K., Funkhouser, H., Lyman, G., & Huff, E.M. (1976). NASA Aviation Safety Reporting System. (Technical Report TM-X-3445). Moffett Field, CA: NASA Ames Research Center.

Billings, C.E., & Woods, D.D. (1994). Concerns about adaptive automation in aviation systems. In M. Mouloua, & R. Parasuraman (Eds.), Human performance in automated systems: Current research and trends (pp. 264-269). Hillsdale, NJ: Lawrence Erlbaum Associates.

Christ, R.E., Hill, S.G., Ayers, J.C., Iavecchia, H.M., Zaklad, A.L., & Bittner, A.C. (1993). Application and validation of workload assessment techniques. Technical Report 974. U.S. Army Research Institute for the Behavioral Sciences, Alexandria, VA.

Clamann, M.P., Wright, M.C., & Kaber, D.B. (2002). Comparison of performance effects of adaptive automation to various stages of human-machine system information processing. In Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting. Santa Monica, CA: Human Factors and Ergonomics Society.

Comstock, J.R., & Arnegard, R.J. (1992). The Multi-Attribute Task Battery for Human Operator Workload and Strategic Behavior Research (Technical Memorandum No. 104174). Hampton, VA: NASA Langley Research Center.

Crocoll, W.M., & Coury B.G. (1990). Status or recommendation: Selecting the type of information for decision aiding. In Proceedings of the Human Factors and Ergonomics Society 34th Annual Meeting (pp. 1524-1528). Santa Monica, CA: Human Factors and Ergonomics Society.

Davison, H.J., & Wickens, C.D. (2001). Rotorcraft hazard cueing: The effects on attention and trust. In Proceedings of the 11th International Symposium on Aviation Psychology (pp. 1-6). Columbus, OH: The Ohio State University.

Doton, L. (1996). Integrating technology to reduce fratricide. Acquisition Review Quarterly, (Winter Issue), 1-18.

Duley, J.A., Westerman, S., Molloy, R. & Parasuraman, R. (1997). Effects of display superimposition on monitoring of automation. In Proceedings of the 9th International Symposium on Aviation Psychology (pp. 322-328). Columbus, OH: Ohio State University.

Dzindolet, M.T., Pierce, L. Pomranky, R. Peterson, S., & Beck, H. (2001). Automation reliance on a combat identification system. In Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting (pp. 532-536). Santa Monica, CA: Human Factors and Ergonomics Society.

Farrell, S., & Lewandowsky, S. (2000). A connectionist model of complacency and adaptive recovery under automation. Journal of Experimental Psychology: Learning, Memory and Cognition, (26), 395-410.

Fitts, P. (1951). Human engineering for an effective air navigation and traffic control system. National Research Council, Washington, DC: Author.

Galster, S.M., Bolia, R.S, & Parasuraman, R. (2002). Effects of information automation and decision-aiding cueing on action implementation in a visual search task. In Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting (pp. 438-442). Santa Monica, CA: Human Factors and Ergonomics Society.

Galster, S.M., Bolia, R.S., Roe, M.M., & Parasuraman, R. (2001). Effects of automated cueing on decision implementation in a visual search task. In Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting (pp. 321-325). Santa Monica, CA: Human Factors and Ergonomics Society.

Hancock, P.A., & Chignell, M.H. (1987). Adaptive control in human-machine systems. In P.A. Hancock (Ed.), Human factors psychology (pp. 305-345). North Holland: Elseveir.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, In P.A. Hancock and N. Meshkati (Eds.) Human mental workload. Amsterdam: Elsveir

Hendy, K.C., Hamilton, K.M., & LandryL.N. (1993). Measuring subjective workload: When is one scale better than many? Human Factors, 35, 579-601.

Howell, W.C., Johnston, W.A., & Goldstein, I.L. (1966). Complex monitoring and its relation to the classical problem of vigilance. Organizational Behavior and Human Performance, (1), 129-150

Hurst, K., & Hurst, L. (1982). Pilot error: The human factors. New York: Aronson.

Jordan, N. (1963). Allocation of functions between man and machines in automated systems. Journal of Applied Psychology, 47, 161-165.

Kennedy, R.S., Lane, N.E., Berbaum, K.S. & Lilienthal, M.G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. The International Journal of Aviation Psychology, 3(3), 203- 220.

Kessel, C.J., & Wickens, C.D. (1982). The transfer of failure-detection skills between monitoring and controlling dynamics. Human Factors, 24, 49-60.

Knapp, R.K., & Vardaman, J.J. (1991). Response to an automated function failure cue: an operational measure of complacency. In Proceedings of the Human Factors Society, 35th Annual Conference, San Francisco, CA: Human Factors Society.

Lee, J.D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35, 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. International Journal of Human-Computer Studies, 40, 153-184.

Lee, J.D., & Sanquist, T.F. (1996). Maritime Automation. In R. Parasuraman and M. Mouloua (Eds.), Automation and Human Performance: Theory and Applications (pp. 365-384). Hillsdale, NJ: Erlbaum.

Lorenz, B., Di Nocera, F., & Parasuraman, R. (2002). Display integration enhances information and decision making in automated fault management in a simulated spaceflight micro-world. In Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society (pp. 31-35). Santa Monica, CA: Human Factors and Ergonomics Society.

Masalonis, A.J. (2000). Effects of situation-specific reliability on trust and usage of automated decision aids. Unpublished doctoral dissertation. Washington, DC: The Catholic University of America.

McGarry, K., Rovira, E. & Parasuraman, R. (in press). Effects of task duration and type of automation support on human performance and stress in a simulated battlefield engagement task. In Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society. Santa Monica, CA: Human Factors and Ergonomics Society.

Merlo, J.L., Wickens, C.D., & Yeh, M. (2000). Effect of reliability on cue effectiveness and display signaling. In Proceedings of the 4th Annual Army Federated Laboratory Symposium (pp. 27-31). College Park, MD: Army Research Federated Laboratory Consortium.

Metzger, U., & Parasuraman, R. (2001). Conflict detection aids for air traffic controllers in free flight: Effects of reliable and failure modes on performance and eye movements. In Proceedings of the 11th International Symposium on Aviation Psychology (pp. 1-5). Columbus, OH: Ohio State University.

Miller, C.A., & Parasuraman, R. (in press). Beyond levels of automation: An architecture for more flexible human-automation collaboration. In Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society. Santa Monica, CA: Human Factors and Ergonomics Society.

Middendorf, M.S., Galster, S.M., & Brown, R.D. (2003). Creating a modular experimental simulation environment using ModSAF. In Proceedings of the 12th International Symposium on Aviation Psychology (pp. 810-814). Dayton, OH: The Wright State University.

Molloy, R., & Parasuraman, R. (1994). Automation-induced monitoring inefficiency: The role of display integration and redundant color coding. In M. Mouloua and R. Parasuraman (Eds.), Human Performance in Automated Systems: Current Research and Trends (pp. 224-228). Hillsdale, NJ: Erlbaum.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. Human Factors, 38(2), 311-322.

Mosier, K.L., Skitka, L.J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. The International Journal of Aviation Psychology, 8(1), 47-63.

Mosier, K.L., Skitka, L.J., & Korte, K.J. (1994). Cognitive and social psychological issues in flight crew/automation interaction. In M. Mouloua and R. Parasuraman (eds.), Human performance in automated systems: Current research and trends (pp. 191-197). Hillsdale, NJ: Erlbaum.

Mouloua, M., & Parasuraman, R. (1995). Aging and cognitive vigilance: Effects of spatial uncertainty and event rate. Experimental Aging Research, 21, 17-32.

Muir, B.M. (1988). Trust between humans and machines, and the design of decision aids. In E. Hollnagel, G. Mancini, and D.D. Woods (Eds.), Cognitive engineering in complex dynamic worlds (pp. 71-83). London: Academic Press.

Muthard, E.K., & Wickens, C.D. (2001). Change detection and the confirmation bias in aviation route planning (Technical Report ARL-01-18/NASA-01-9). Savoy, IL: University of Illinois, Aviation Research Lab.

National Transportation Safety Board (1973). Eastern Airlines L-1011, Miami, Florida, December 29, 1972 (Report No. NTSB-AAR-73-14). Washington, DC: Author.

National Transportation Safety Board (1986). China Airlines Boeing 747-SP, N4522V, 300 nautical miles northwest of San Francisco, California 1985 (Report No. NTSB-AAR-86-03). Washington, DC: Author.

Nygren, T.E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. Human Factors, 33, 17-31.

Parasuraman, R. (1993). Effects of adaptive function allocation on human performance. In D.J. Garland and J.A. Wise (Eds.), Human Factors and Advanced Aviation Technologies (147-157). Daytona, FL: Embry-Riddle Aeronautical University Press.

Parasuraman, R., Bahri, T., Deaton, J., Morrison, J., & Barnes, M. (1992). Theory and design of adaptive automation in aviation systems. (Technical Report No. NAWCADWAR-92033-60). Warminster, PA: Naval Air Warfare Center.

Parasuraman, R., Bahri, T., Molloy, R. (1991). Adaptive automation and human performance: I. Multi-task performance characteristics (Tech. Report CSL-N91-1). Washington, DC: The Catholic University of America, Cognitive Science Laboratory.

Parasuraman, R., Molloy, R., & Singh, I.L. (1993). Performance consequences of automation induced "complacency". International Journal of Aviation Psychology, 3, 1-23.

Parasuraman, R., & Mouloua, M. (1996). Automation and human performance: Theory and applications. Mahwah, NJ: Erlbaum.

Parasuraman, R., Mouloua, M., & Hilburn, B. (1999). Adaptive aiding and adaptive task allocation enhance human-machine interaction. In M. Scerbo and M. Mouloua (Eds.), Automation technology and human performance: Current research and trends (pp. 119-123). Mahwah, NJ: Erlbaum.

Parasuraman, R., Mouloua, M., Molloy, R. (1994). Monitoring automation failures in human-machine systems. In M. Mouloua & R. Parasuraman (Eds.), Human performance in automated systems: Current research and trends (pp. 45-49). Hillsdale, NJ: Lawrence Erlbaum Associates.

Parasuraman, R., Mouloua, M., Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. Human Factors, 38, 665-679.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39, 230-253.

Parasuraman, R., Sheridan, T.B., Wickens, C.D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 30, 286-297.

Park, K.S. (1997). Human Error. In G. Salvendy (Ed.), Handbook of human factors and ergonomics, second edition (pp. 150-173). New York, NY: Wiley & Sons.

Perrow, C. (1999). Normal accidents. Princeton, NJ: Princeton University Press.

Reason, J.T. (1990). Human error. Cambridge, United Kingdom: Cambridge University Press.

Rouse, W.B., & Rouse, S.H. (1983). A framework for research on adaptive decision aids (Technical Report AFAMRL-TR-83-082). Wright-Patterson Air Force Base, OH: Air Force Aerospace Medical Research Laboratory.

Rovira, E., McGarry, K., & Parasuraman, R. (2002). Effects of unreliable automation on decision making in command and control. In Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society (pp. 428-432). Santa Monica, CA: Human Factors and Ergonomics Society.

Rovira, E., Zinni, M., & Parasuraman, R. (2002). Effects of information and decision automation on multi-task performance. In Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society (pp. 327-331). Santa Monica, CA. Human Factors and Ergonomics Society.

Sarter, N., & Schroeder, B. K. (2001). Supporting decision-making and action selection under time pressure and uncertainty: The case of in-flight icing. Human Factors, 43 (4), 573-583.

Sarter, N.B., & Woods, D.D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Human Factors, 37, 5-19.

Sarter, N.B., Woods. D.D., & Billings, C.E. (1997). Automation Surprises. In G. Salvendy (Ed.), Handbook of human factors and ergonomics (2nd ed., pp. 1926-1943). New York: Wiley.

Sheridan, T. B. (1980). Computer control and human alienation. Technology Review, 10, 61-73.

Sheridan, T.B. (1988). Trustworthiness of command and control systems. In Proceedings of the International Federation of Automatic Control Conference on Man-Machine Systems (pp. 427-431). Elmsford, NY: Pergamon.

Sheridan, T.B. (2002). Humans and automation: System design and research issues. Hoboken, NJ: Wiley and Sons.

Sheridan, T.B., & Verplank, W.L. (1978). Human and computer control of undersea tele-operators (Man-Machine Systems Laboratory Report). Cambridge, MA: MIT Flight Technology Laboratory.

Singh, I.L., Molloy, R., & Parasuraman, R. (1993). Individual differences in monitoring failures of automation. Journal of General Psychology, 120(3), 357-373.

Singh, I.L., Molloy, R., & Parasuraman, R. (1997). Automation-induced monitoring inefficiency: role of display location. International Journal of Human-Computer Studies (46), 17-30.

Stein, K.J. (1983, October 3). Human factors analyzed in 007 navigation error. Aviation Week & Space Technology, 165-167.

Sumwalt, R.L., Morrison, R., Watson, A., & Taube, E. (1997). What ASRA data tell about inadequate flight crew monitoring. In Proceedings of the 9th International Symposium on Aviation Psychology (pp. 977-982). Columbus, OH: Ohio State University.

Taylor, R.M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In Situational Awareness in Aerospace Operations (AGARD-CP-478, pp. 3-1 and 3-17). Neuilly Sur Seine, France: NATO-AGARD.

Thackray, R.I., & Touchstone, R.M. (1989). Detection efficiency on an air traffic control monitoring task with and without computer aiding. Aviation, Space, and Environmental Medicine, 60, 744-748.

Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12, 97-136.

Wickens, C.D. (1994). Designing for situation awareness and trust in automation. In Proceedings of the IFAC Conference on Integrated Systems Engineering. Baden-Baden, Germany: International Federation of Automatic Control.

Wickens, C.D., & Carswell C.M. (1997). Information processing. In G. Salvendy (Ed.), Handbook of human factors and ergonomics, second edition (pp. 89-129). New York, NY: Wiley & Sons

Wickens, C.D., Conejo, R., & Gempler, K. (1999). Unreliable automated attention cueing for air-ground targeting and traffic maneuvering. In Proceedings of the Human Factors and Ergonomics Society the 34th Annual Meeting (pp. 21-25). Santa Monica, CA: Human Factors and Ergonomics Society.

Wickens, C.D., & Dixon, S. (2002). Workload demands of remotely piloted vehicle supervision and control: (I) Single vehicle performance (ARL-02-10/MAD-02-1). Savoy, IL: University of Illinois, Aviation Research Lab.

Wickens, C. D., Mavor, A. S., & McGee, J. P. (1997). Flight to the future: Human factors in air traffic control. Washington, DC: National Academy Press.

Wickens, C.D., & Xu, X. (2002). Automation trust, reliability and attention HMI 02-03 (Technical Report No. AHFD-02-14/MAAD-02-2). Savoy, IL: University of Illinois, Aviation Research Lab.

Wiener, E.L. (1981). Complacency: Is the term useful for air safety? In Proceedings of the 26th Corporate Aviation Safety Seminar (pp. 116-125). Denver, CO: Flight Safety Foundation.

Wiener, E.L. (1988). Cockpit automation. In E.L. Wiener and D.C. Nagel, (Eds.), Human Factors in Aviation. (pp. 433-461). San Diego, CA: Academic Press.

Wiener, E.L., & Curry, R.E. (1980). Flight-deck automation: Promises and problems. Ergonomics, 23, 995-1011.

Wiener, E.L., & Curry, R.E., & Faustina, M.L. (1984). Vigilance and task load: In search of the inverted U. Human Factors, 26, 215-222.

Woods. D.D. (1996). Decomposing automation: Apparent simplicity, real complexity. In R. Parasuraman and M. Mouloua (Eds.) Automation and Human Performance: Theory and Applications (pp. 1-17). Mahwah, NJ: Erlbaum.

Yeh, M., & Wickens, C.D. (2001). Display signaling in augmented reality: The effects of cue reliability and image realism on attention allocation and trust calibration. Human Factors, 43(3), 355-365.

Yeh, M., Wickens, C.D., & Seagull, F.J. (1999). Target cueing in visual search: The effects of conformality and display location on the allocation of visual attention. Human Factors, (41), 524-542.